华为职业认证通过者权益

通过**任一项**华为职业认证，您即可在华为在线学习网站(*http://learning.huawei.com/cn*) 享有如下特权：

- **1、华为E-learning 课程学习**
  - **内容**：**所有华为职业认证E-Learning课程**，扩展您在其他技术领域的技术知识
  - **方式**：请提交您的 "华为账号" 和注册账号的 "email地址" 到 *Learning@huawei.com 申请权限。*
- **2、华为培训教材下载**
  - **内容：华为职业认证培训教材+华为产品技术培训教材，**覆盖企业网络、存储、安全等诸多领域
  - **方式**：登录*华为在线学习网站*，进入 "*华为培训->面授培训*"，在具体课程页面即可下载教材。
- **3、华为在线公开课(LVC)优先参与**
  - **内容**：企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师授课，开班人数有限
  - **方式**：开班计划及参与方式请详见LVC排期：
    http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_16
- **4、学习工具 eNSP**
  - *eNSP (Enterprise Network Simulation Platform)*, 是由华为提供的免费的、可扩展的、图形化网络仿真工具。主要对企业网路由器和交换机进行硬件模拟，完美呈现真实设备实景；同时也支持大型网络模拟，让大家在没有真实设备的情况下也能够进行实验测试。
- 另外，华为建立了知识分享平台 华为认证论坛。您可以在线与华为技术专家交流技术，与其他考生分享考试经验，一起学习华为产品技术。（http://support.huawei.com/ecommunity/bbs/list_2247.html ）

华为认证系列教程

HCNP-Storage CBDS

构建大数据解决方案

华为技术有限公司

华为认证系列教程

HCNP-Storage CBDS 构建大数据解决方案

第2.0版本

华为认证系统介绍

依托华为公司雄厚的技术实力和专业的培训体系，华为认证考虑到不同客户对ICT技术不同层次的需求，致力于为客户提供实战性、专业化的技术认证。 根据ICT技术的特点和客户不同层次的需求，华为认证为客户提供面向十三个方向的四级认证体系。

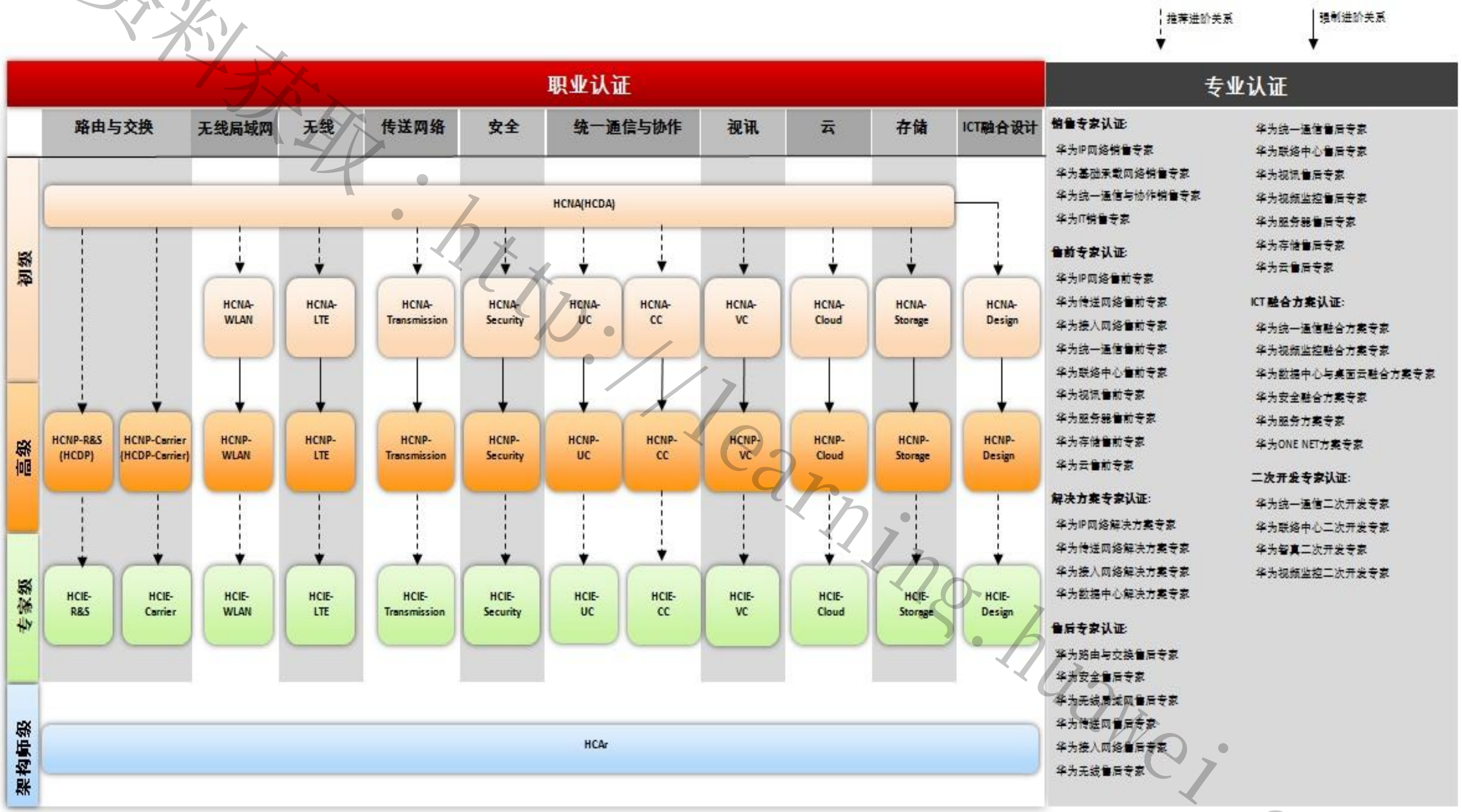
HCNA-Storage（Huawei Certified Network Associate-Storage，华为网络存储工程师认证）主要面向网络存储维护工程师，以及准备参加HCNA-Storage认证考试的人员；希望掌握SAN存储系统与网络基本原理和华为SAN存储阵系统管理、部署与维护能力的人员。HCNA认证在内容上涵盖存储基础知识、RAID技术与应用、存储网络技术与应用、华为存储产品与解决方案、存储系统管理和基本配置、存储主机连接与多路径配置、SAN网络与存储系统日常维护。

HCNP-Storage (Huawei Certified Network Professional-Storage，华为认证网络存储资深工程师)主要面向企业级网络存储维护工程师、专家工程师以及希望系统深入掌握SAN存储、统一存储、数据保护技术与部署的人员。

HCNP-Storage包括CUSN（Constructing Unifying Storage Network，构建统一存储网络）、CBDS（Constructing Big Data Solution 构建大数据解决方案）、CDPS（Constructing Data Protection System 构建数据保护系统）三个部分。内容上涵盖SAN、NAS统一存储原理、架构和组件，存储数据处理与通信协议（SCSI、FC、iSCSI）原理及应用，存储系统数据可靠性与业务连续性保障技术存储与主流OS平台连接与应用，存储网络冗余技术及应用，SAN、集群NAS 网络规划与方案部署，虚拟快照、LUN Copy、复制的原理和部署，网络存储虚拟化技术及应用，存储虚拟化系统原理、部署和异构资源管理，备份网络及备份恢复技术及应用，华为数据保护方案构建、部署与管理，华为数据容灾方案及典型应用场景，华为存储系统、网络、方案的故障诊断与处理方法。

HCIE-Storage（Huawei Certified Internetwork Expert--Storage，华为认证互联网络存储专家）旨在培养能够熟练掌握各种存储技术；精通 IT 存储方案设计、部署和运维管理以及诊断和故障排除。

华为认证协助您打开行业之窗，开启改变之门，屹立在ICT世界的潮头浪尖!


推荐进阶关系
强制进阶关系
职业认证
专业认证
路由与交换
无线局域网
无线
传送网络
安全
统一通信与协作
视讯
云
存储
ICT融合设计
初级
高级
专家级
架构师级
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr
销售专家认证:
华为IP网络销售专家
华为基础承载网络销售专家
华为统一通信与协作销售专家
华为IT销售专家
售前专家认证:
华为IP网络售前专家
华为传送网络售前专家
华为接入网络售前专家
华为统一通信售前专家
华为联络中心售前专家
华为视讯售前专家
华为服务器售前专家
华为存储售前专家
华为云售前专家
解决方案专家认证:
华为IP网络解决方案专家
华为传送网络解决方案专家
华为接入网络解决方案专家
华为数据中心解决方案专家
售后专家认证:
华为路由与交换售后专家
华为安全售后专家
华为无线测试网售后专家
华为传送网售后专家
华为接入网络售后专家
华为无线售后专家
华为统一通信售后专家
华为联络中心售后专家
华为视讯售后专家
华为视频监控售后专家
华为服务器售后专家
华为存储售后专家
华为云售后专家
ICT融合方案认证:
华为统一通信融合方案专家
华为视频监控融合方案专家
华为数据中心与桌面云融合方案专家
华为安全融合方案专家
华为服务方案专家
华为ONE NET方案专家
二次开发专家认证:
华为统一通信二次开发专家
华为联络中心二次开发专家
华为智真二次开发专家
华为视频监控二次开发专家

前言

简介

HCNP-Storage 认证定位于 IT 领域信息存储高级工程师或存储方案专家能力构建。

HCNP-Storage 认证培训教程，适用于华为认证网络存储资深工程师以及准备参加 HCNP-Storage 认证考试的人员，通过 HCNP-Storage 认证，将证明您对信息存储系统、网络有全面深入的理解，掌握存储系统和网络（SAN、NAS、灾备系统）的通用技术，并具备独立完成信息存储、数据保护等综合解决方案的部署、运维和管理。

内容描述

本书是HCNP-Storage-CBDS（Huawei Certified Network Professional - Constructing Big Data Solution华为认证网络存储资深工程师－构建大数据解决方案），培训用于指导学员学习HCNP-Storage-CBDS认证考试（H13-622）相关内容。内容覆盖：大数据概论、大数据的产生和发展，大数据存储技术、分布式文件系统动态以及动态分级存储，大数据存储系统部署与管理，N8500的原理特性安装配置与集群，非线性编辑业务（媒资场景），大数据存储系统故障诊断，OceanStor9000和OceanStorN8500的故障分类定义诊断，大数据存储解决方案规划设计。

Module 1 大数据概论：大数据概念与特征，大数据产生与发展，大数据行业实践。

Module 2 大数据存储技术：分布式数据可靠技术，分布式文件系统，动态分级存储，负载均衡，配额管理技术，高速互联与接口协议，数据安全，集群技术，文件系统镜像，文件系统快照，文件系统复制。

Module 3 集群 NAS 系统部署与管理：集群 NAS N8500 简介，集群 NAS N8500 原理与特性，集群 NAS N8500 安装与配置，集群 NAS N8500 日常运维。

Module 4 大数据存储系统部署与管理：大数据存储简介，大数据存储原理与特性，大数据存储系统安装与配置，大数据存储日常运维

Module 5 集群 NAS 存储系统故障处理：N8500 故障处理的原则、方法，N8500 故障处理的流程 ，N8500 分类故障以及典型问题处理方法，N8500 故障处理案例。

Module 6 大数据存储系统故障诊断：OceanStor9000 故障分类和定义，OceanStor9000 故障诊断流程，OceanStor9000 典型故障诊断

Module 7 集群 NAS 存储系统规划与设计：N8500 整体规划思路，网络规划，容量与文件系统规划，文件共享业务规划，性能优化。

Module 8 大数据存储规划与设计：OceanStor 9000 前端 GE 后端 10GE 组网部署方案，OceanStor 9000 IB 组网部署方案，OceanStor 9000 GE 组网部署方案。

最终掌握集群、NAS、存储虚拟化技术及应用，并具备独立完成集群 NAS 和存储虚拟化解决方案的规划、部署、运维和管理能力。

读者必备知识背景

本课程为华为认证网络存储高级课程，阅读本书的读者应首先具备以下基本条件：

1、 了解存储技术和 SAN 和 NAS 架构与组成

2、 熟悉 SAN 存储系统配置与管理

3、 熟悉主流 OS 原理和安装部署，熟悉设备管理操作

4、 熟悉文件系统基本知识

5、 熟悉备份、容灾等数据保护概念

6、 有初步的备份项目实施经验

本书常用图标



光纤交换机



以太网交换机



存储系统



主机

目　录

HC1209201 大数据概述

HC1209201
第一章 大数据概述

www.huawei.com

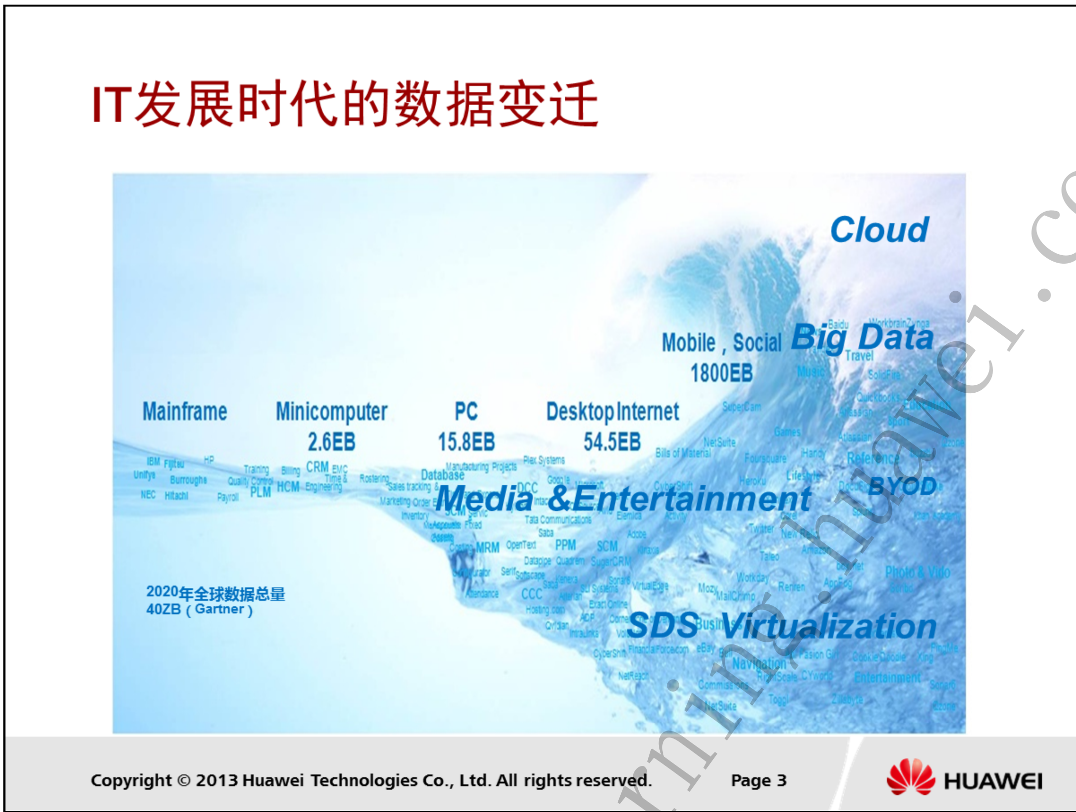
HUAWEI

目标

- 学完本课程后，您将能够:
  - 描述大数据产生、概念和特征；
  - 理解大数据对IT的需求和促进因素
  - 了解大数据方案。

目录

  HUAWEI



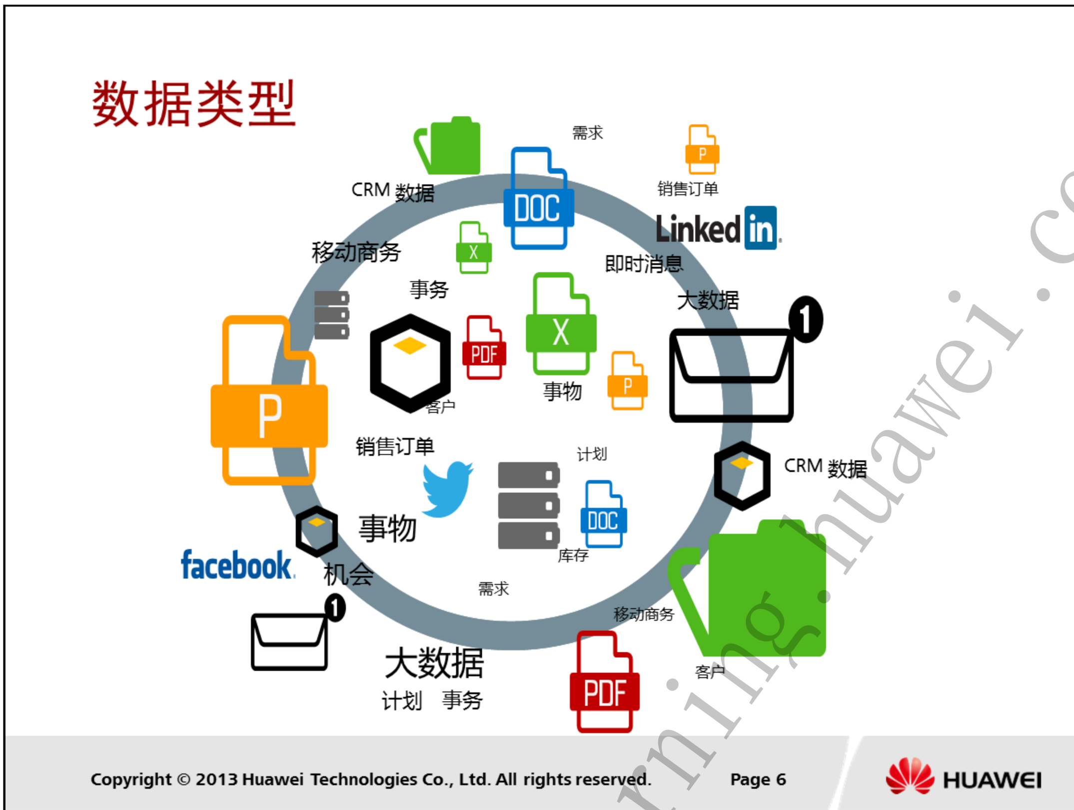
大数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长浪潮的到来，其实，大数据在科学技术领域和军事、金融、通讯等行业并不是新名词，然后其真正进入人们的视野，引起人们的关注却是因为近年来互联网和信息行业的高速发展。



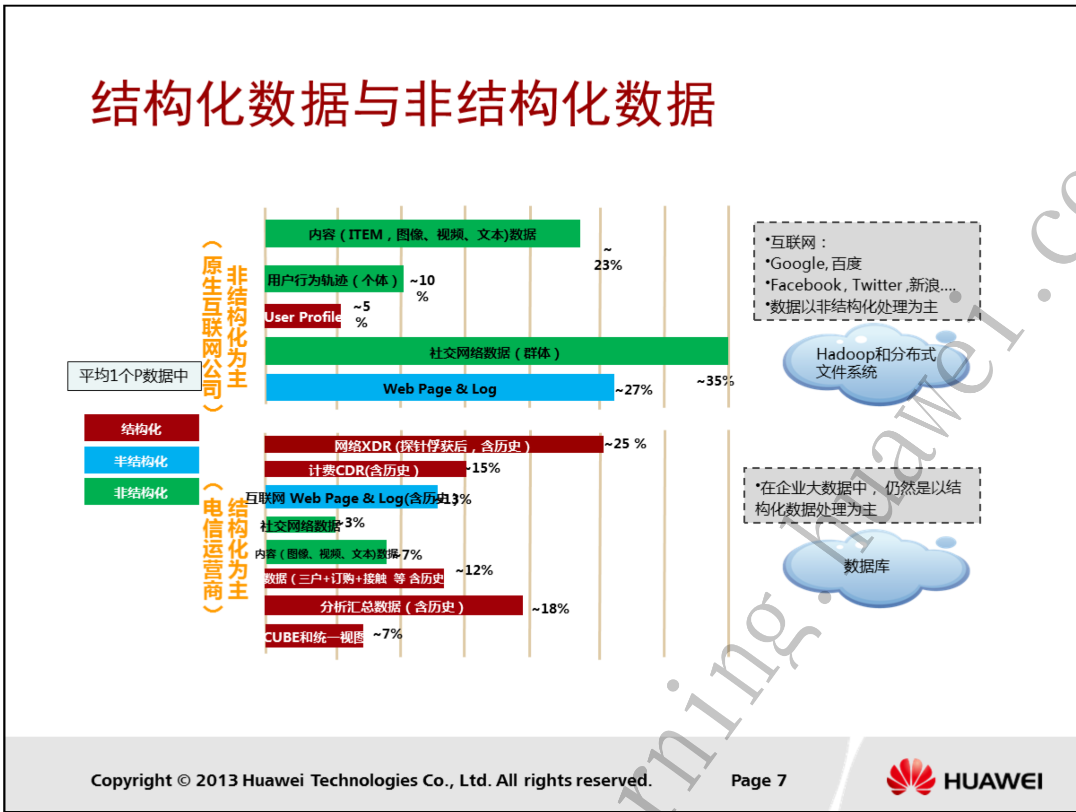
随着人类科学技术的进步，信息的爆炸式增长，人类社会已步入数字化时代。我们每天都会产生和面对大量的数据，同时，我们对数据的需求量越来越大。从以上的数字我们可以看出，数据量级已发生变化，数据价值日益突显。相应的，如何更好的处理这些海量数据，已经成为一个重要的课题。



- Search 检索问题
  - PB级文件系统搜索（包含归类查找和按条件模糊）
  - 百亿记录数据的快速查询
- Efficiency 效率问题
  - 数据的流动性
  - 大量冷数据的归档和节能
  - 重删和压缩
- Store 存储问题
  - 支持Scaleout的可伸缩架构
  - 多协议（标准接口）支撑和多租户架构
  - 数据的可靠性和安全性



在实际应用中，数据因不同的应用场景其表现方式也有所不同，根据在不同应用中数据表现形式的不同，我们可以把数据区分成不同的类型，即数据类型。如：电子商务数据、CRM数据、即时信息、电子邮件等。随着业务应用的丰富多样，数据类型也呈现出其多样性。不同的数据类型，其数据组织的结构、属性、存储方式、分析方法、数据价值等都存在很大的差异，因此，对数据类型的认知和分析将是大数据应用的基础。

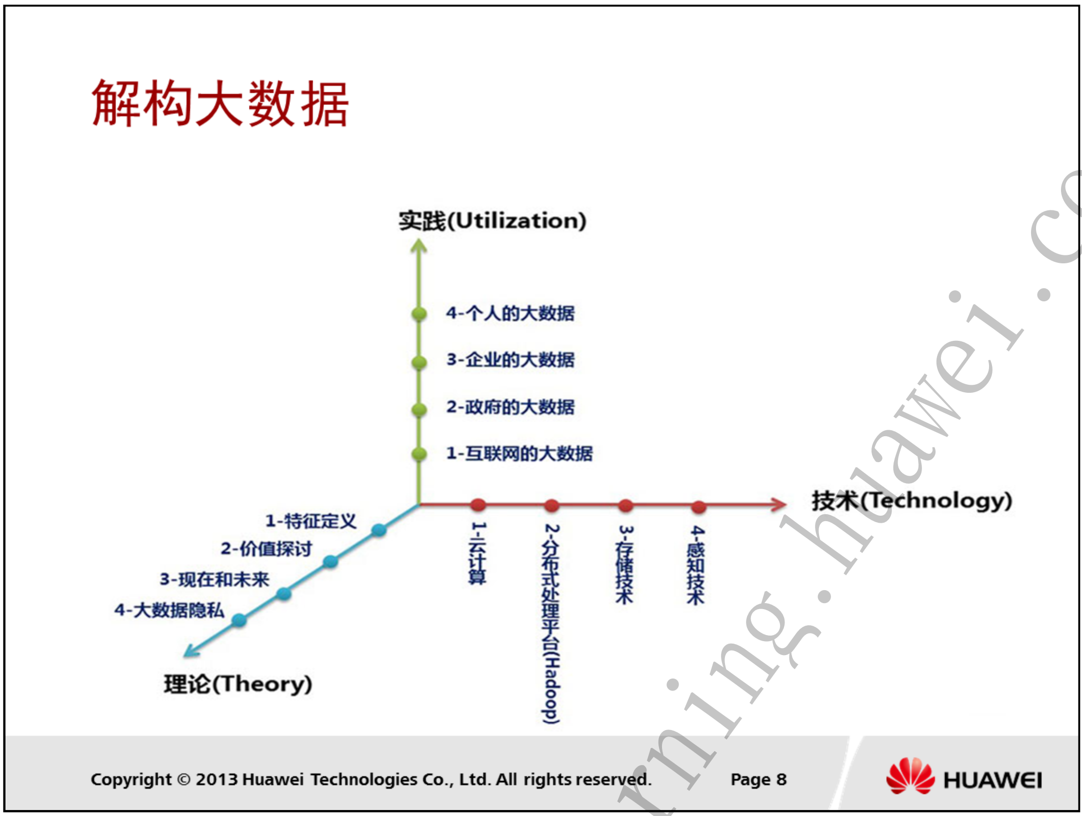


从数据结构特征来分类，主要可以分为：结构化数据、非结构化数据和半结构化数据。

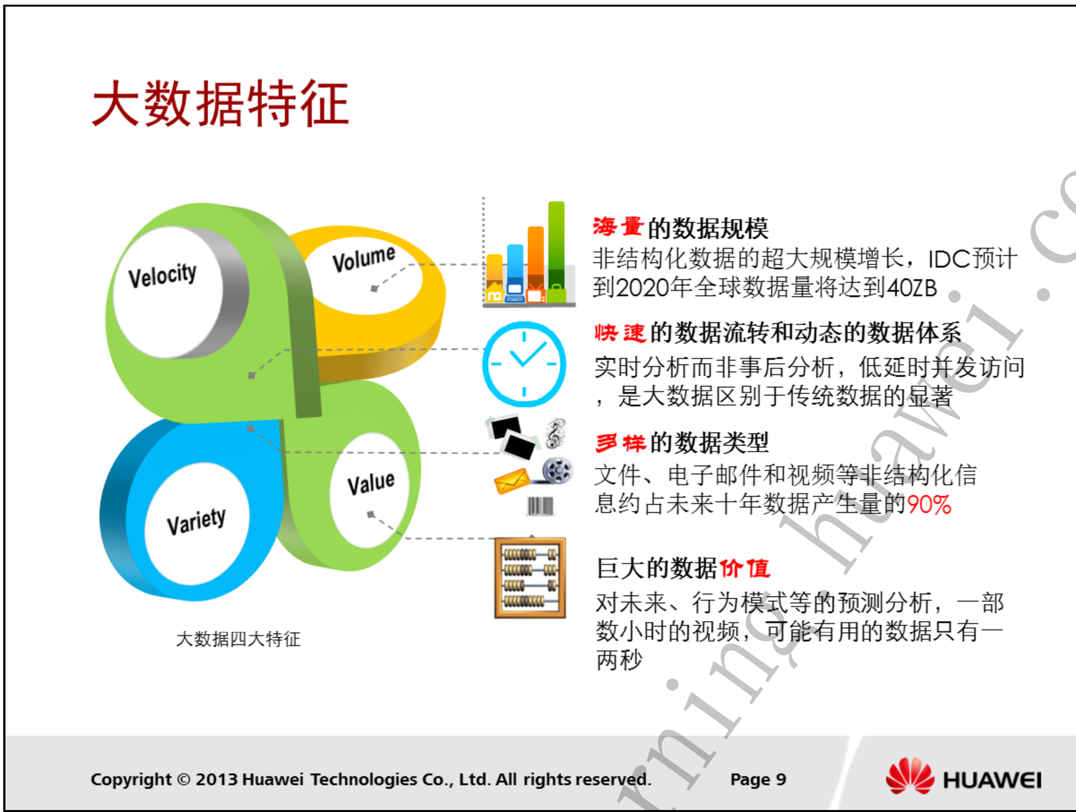
- 结构化数据即指可以用二维表结构来逻辑表达实现的数据，简单来说就是数据库 。比如企业ERP、财务系统、医疗HIS数据库、教育一卡通、政府行政审批等。
- 非结构化数据，即不方便用数据库二维逻辑表来表现的数据。包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等。比如医疗影像系统、教育视频点播、视频监控、国土GIS、设计院、文件服务器（PDM/FTP）、媒体资源管理等。
- 半结构化数据，包括邮件、HTML、报表、资源库等等，典型场景如邮件系统、WEB集群、教学资源库、数据挖掘系统、档案系统等等。

在这里我们主要分析两类大数据应用场景，一类是以互联网为代表，其90%以上数据都是非结构化数据，如Google，百度Facebook，Twitter ,新浪等，数据以非结构化为主。另一类以运营商为代表，其80%以上是结构化为主的大数据。针对这两种大数据代表形式，在业界会有不同的技术处理方式。

大数据的范畴包含结构化和非结构数据，非结构化数据的增长对大数据的促进作用相比较更大、更明显。



我们可以从理论、技术和实践3个维度解剖和理解大数据。每一个维度，都有相应的研究课题、关联技术、方法，在后续的课程学习中将逐步涉及。



大数据的特征按照业界比较认可的说法，归纳为4个“V”（海量Volume，多样Variety，价值Value，快速Velocity），或者说特点有四个层面：

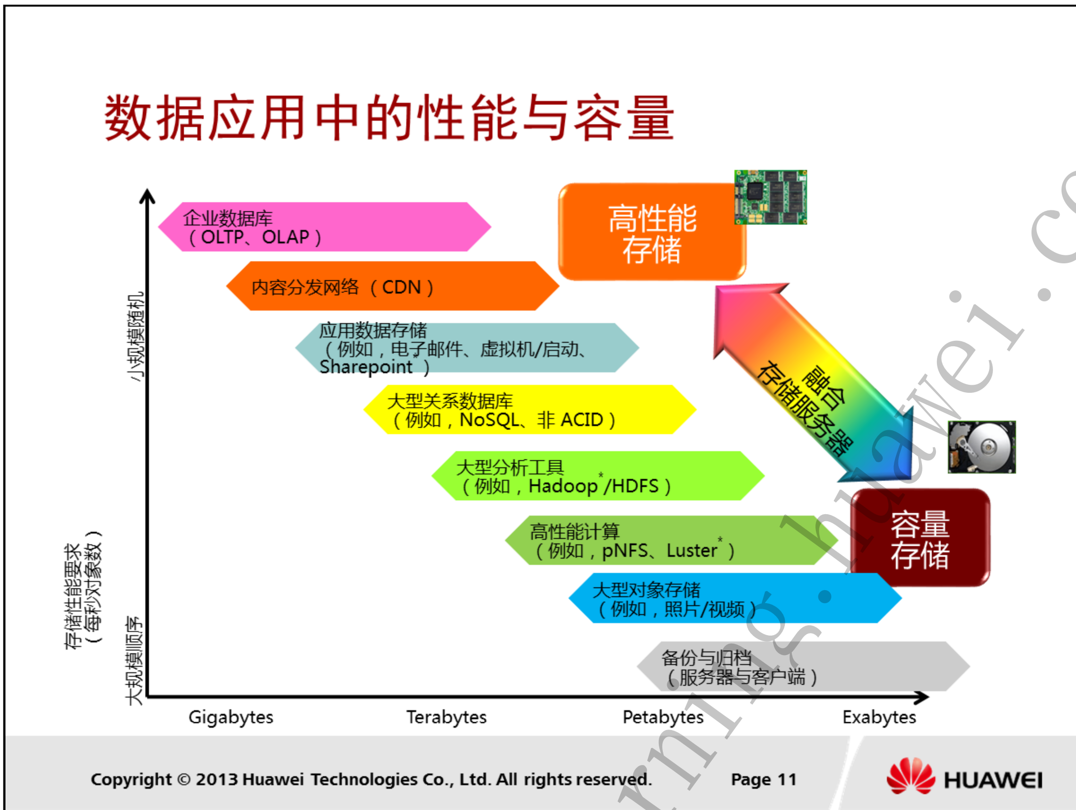
第一，数据体量巨大。大数据的起始计量单位至少是P（1000个T）、E（100万个T）或Z（10亿个T）；

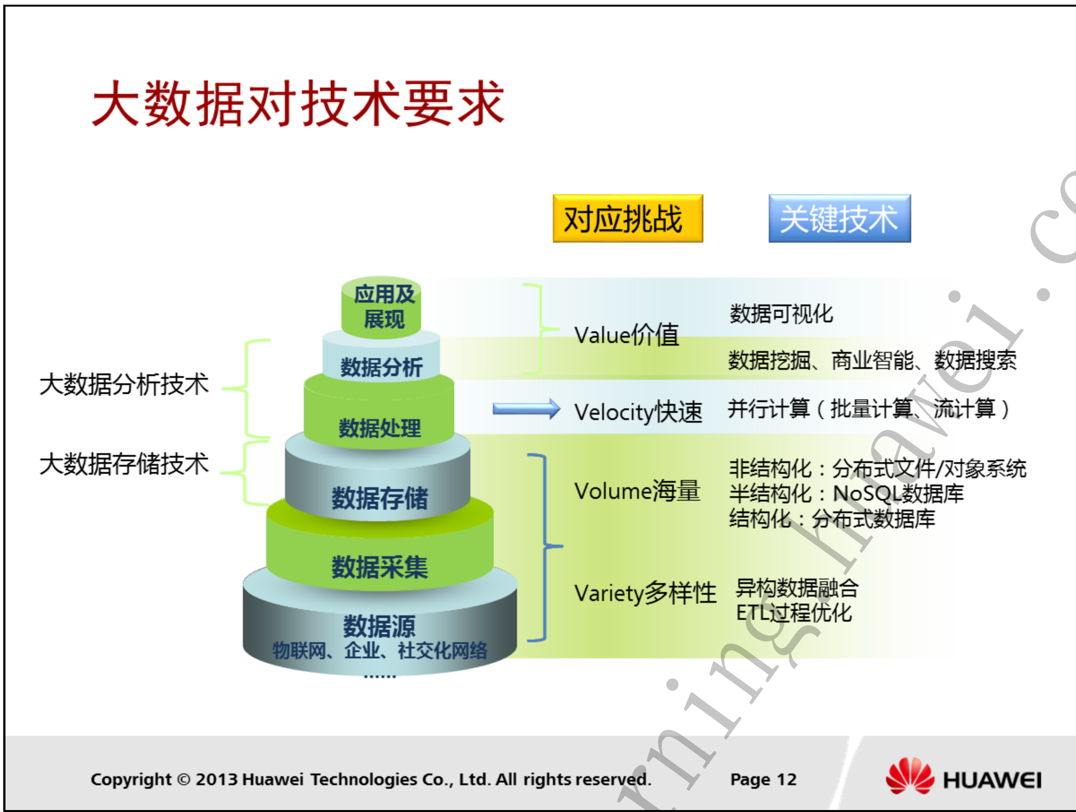
第二，数据类型繁多。比如，网络日志、视频、图片、地理位置信息等等。

第三，价值密度低，商业价值高。

第四，处理速度快。最后这一点也是和传统的数据挖掘技术有着本质的不同。

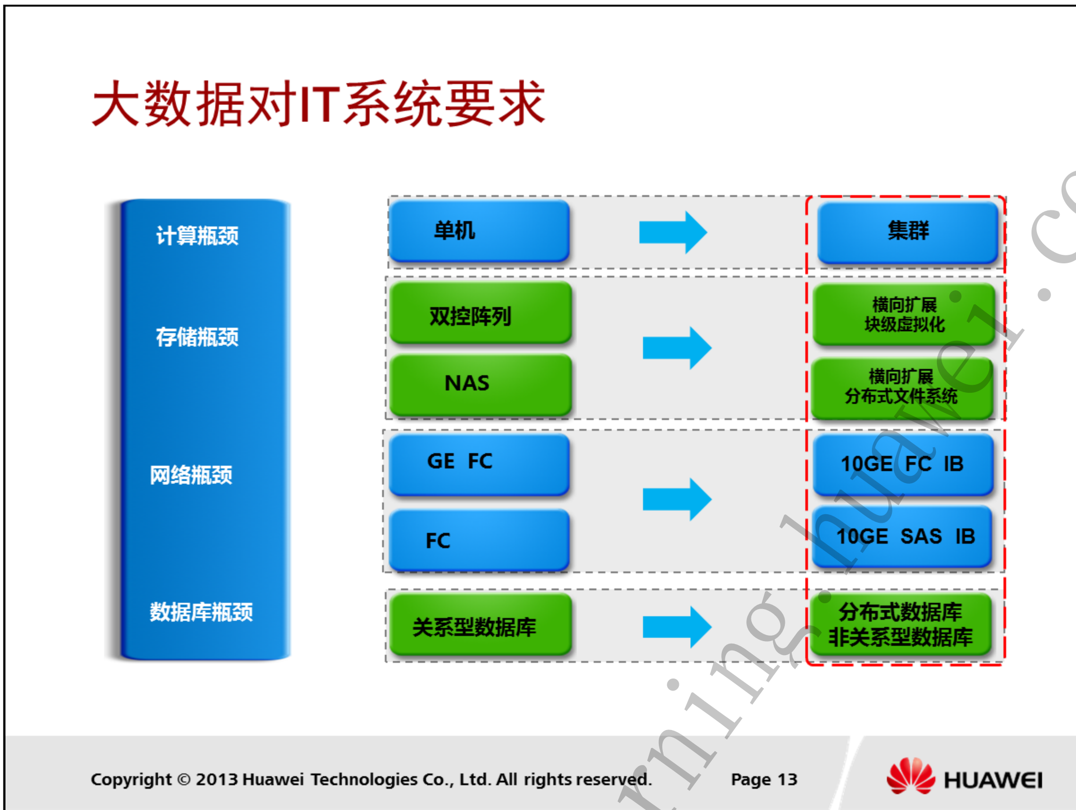
目录


数据应用中的性能与容量
企业数据库
（OLTP、OLAP）
内容分发网络（CDN）
应用数据存储
（例如，电子邮件、虚拟机/启动、Sharepoint）
大型关系数据库
（例如，NoSQL、非 ACID）
大型分析工具
（例如，Hadoop*/HDFS）
高性能计算
（例如，pNFS、Luster*）
大型对象存储
（例如，照片/视频）
备份与归档
（服务器与客户端）
高性能存储
融合存储服务器
容量存储
存储性能要求（每秒对象数）
小规模随机
大规模顺序
Gigabytes
Terabytes
Petabytes
Exabytes
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 11
HUAWEI



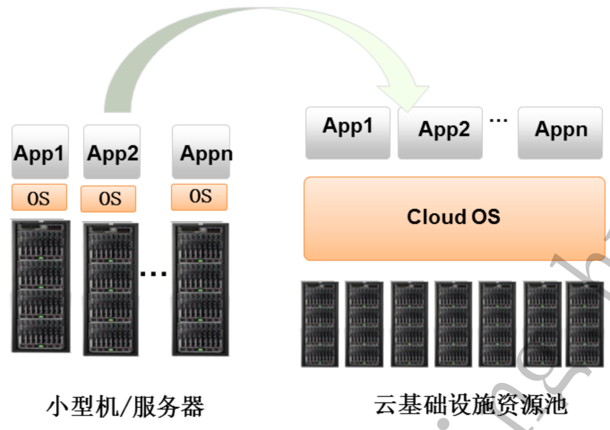
从大数据在容量、数据多样性、处理速度和价值挖掘等四个方面所面临的挑战看，大数据技术涵盖了从数据的海量存储、处理到应用的多方面的技术，包括异构数据源融合、海量分布式文件系统、NoSQL数据库、并行计算框架、实时流数据处理以及数据挖掘、商业智能和数据可视化等。

一个典型的大数据处理系统可分为五层：数据采集、存储、处理、分析、应用及展现等，其技术参考架构如图所示。


大数据对IT系统要求
计算瓶颈
存储瓶颈
网络瓶颈
数据库瓶颈
单机
集群
双控阵列
NAS
横向扩展
块级虚拟化
横向扩展
分布式文件系统
GE FC
FC
10GE FC IB
10GE SAS IB
关系型数据库
分布式数据库
非关系型数据库
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 13
HUAWEI

大数据推动IT架构演进（一）

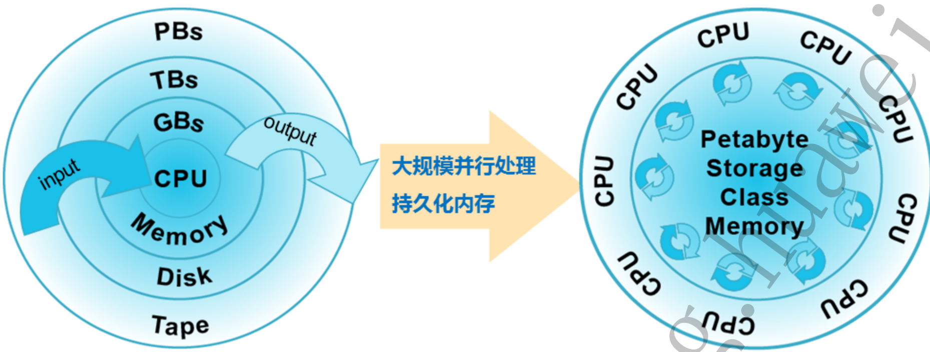
- 应用独占主机→多应用共享大规模资源池



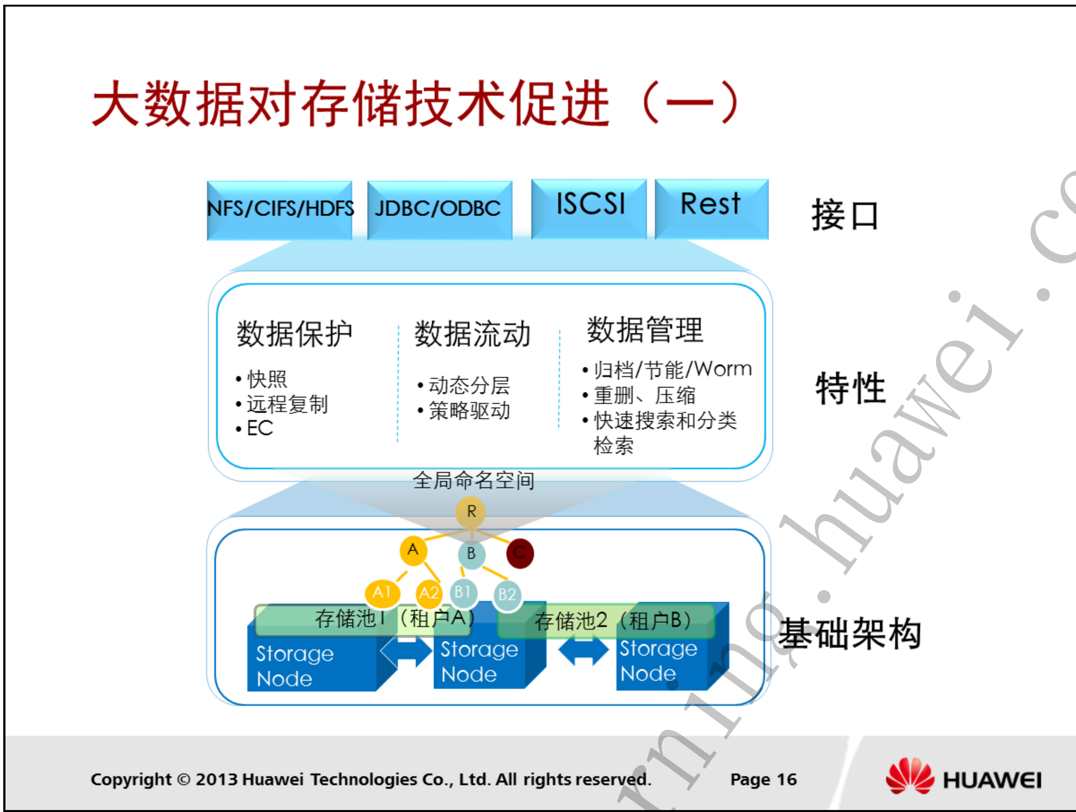
HUAWEI

大数据推动IT架构演进（二）

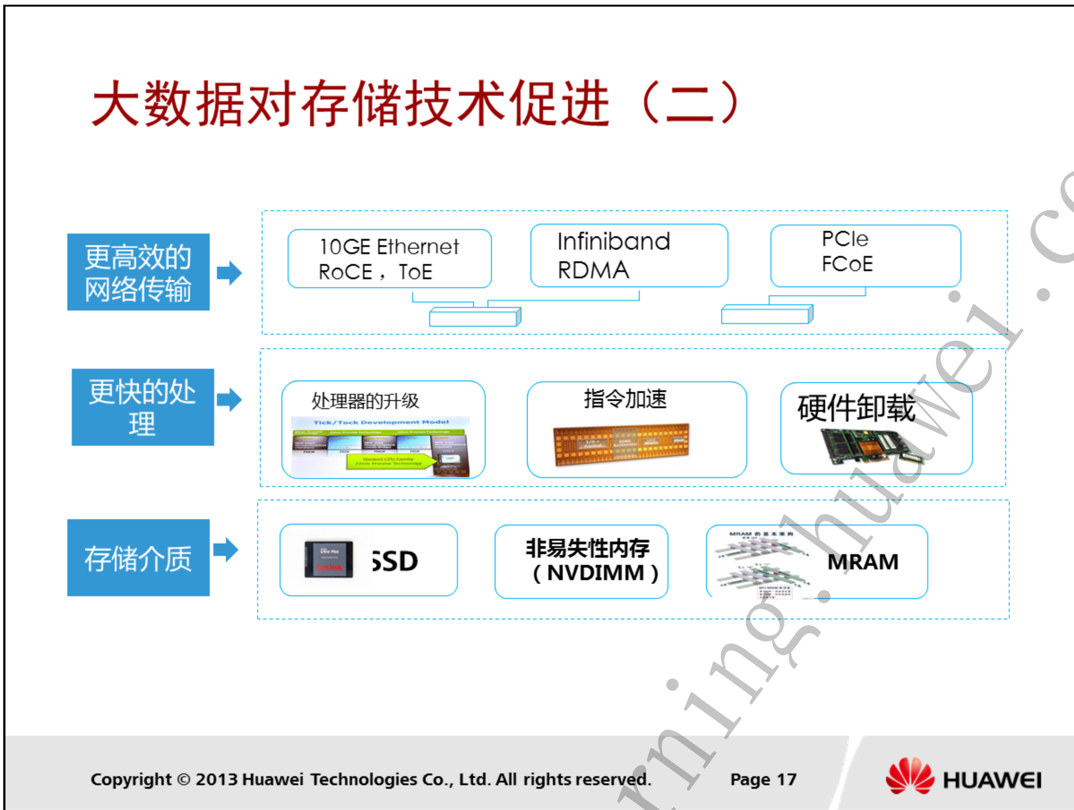
- IT基础设施架构从以计算为中心→以数据为中心

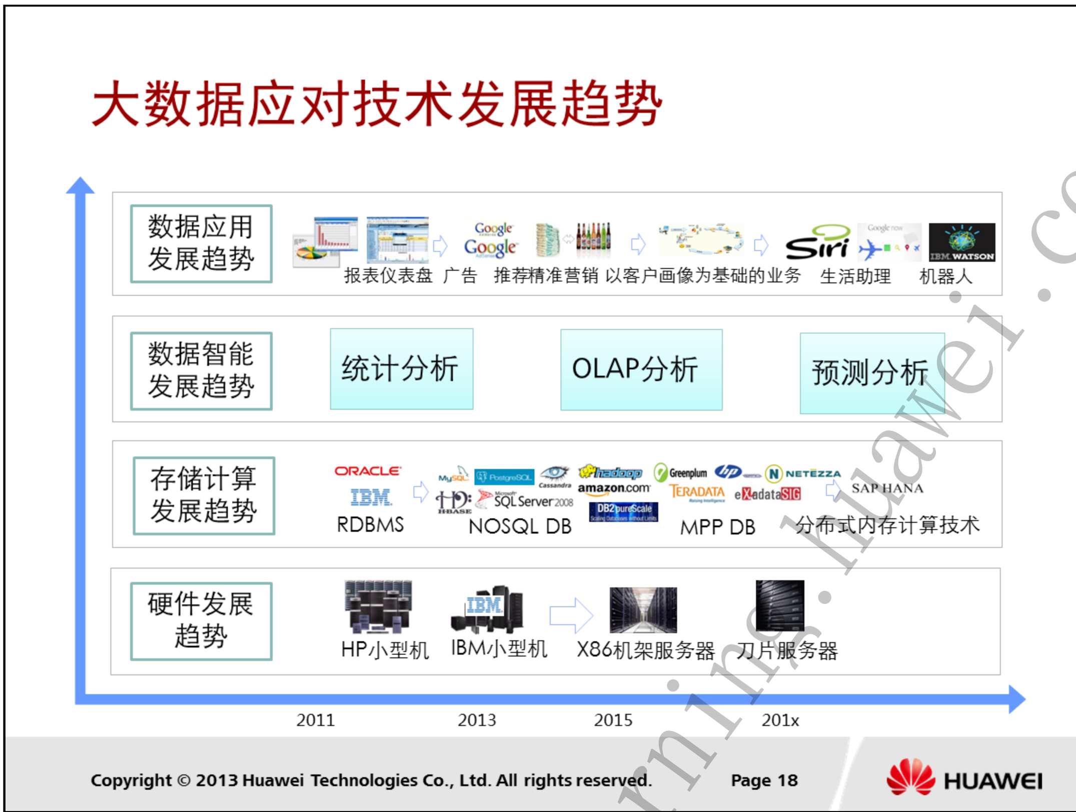


HUAWEI



- 接口
  - NFS/CIFS/Posix接口支持非结构化数据的存取
  - HDFS接口兼容Hadoop分析平台
  - xDBC和SQL语法支持数据库访问方式
  - Rest接口支持对象存储
- 特性
  - 大数据的海量、分布式以及其应用可感知等特点决定了1其保护方式区别于传统Raid和传统的备份。
  - 大数据的数据结构和特征的多样性决定了需要合理的分层和数据流动来保证
  - Capex和Opex
  - 大数据的海量和多样使其管理维护更加困难
- 基础架构
  - 全局命名空间
  - 横向扩展
  - 多租户架构

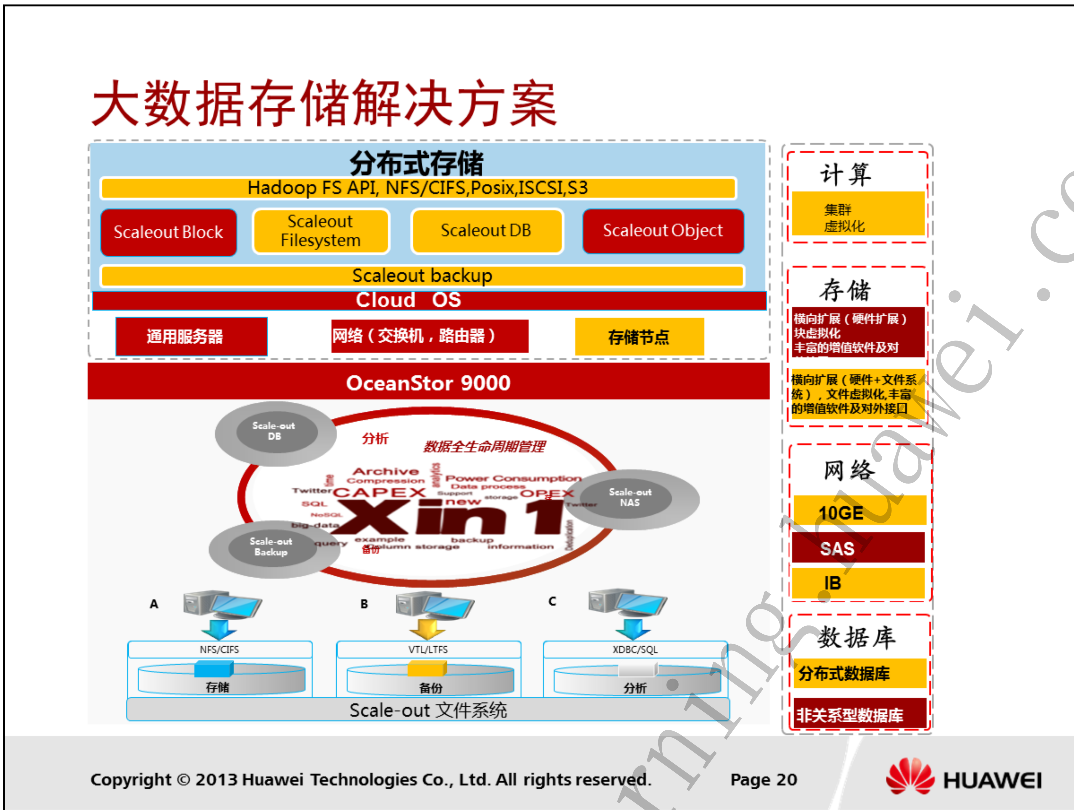

大数据对存储技术促进（二）
更高效的网络传输
10GE Ethernet
RoCE，ToE
Infiniband
RDMA
PCIe
FCoE
更快的处理
处理器的升级
指令加速
硬件卸载
存储介质
SSD
非易失性内存（NVDIMM）
MRAM
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 17
HUAWEI



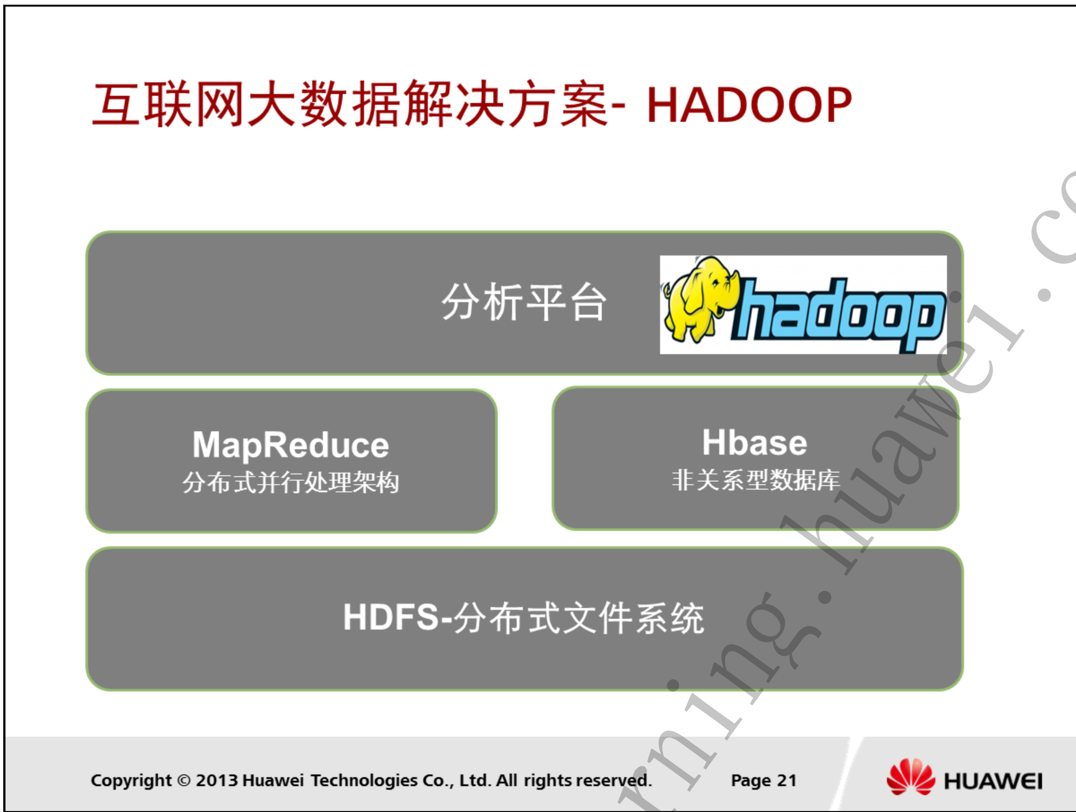
- 数据应用发展趋势
  - 报表仪表盘中展现→广告/推荐/视频质量提升/网络质量提升/精准营销→客户画像为基准的业务
- 数据智能：统计分析→OLAP →数据挖掘与机器学习/实时决策
  - 统计分析：从手工统计，到大规模统计计算，到现在的实时统计，分布式实时统计是当前的主旋律。
  - OLAP分析：预定义多维报表，到T级ROLAP在线分析，演进到分布式的T级，P级多维分析（MOLAP）和Ad hoc分析
  - 预测分析：从基于历史数据的统计预测，发展到机器的挖掘、自学习智能的实时预测能力。
- 存储计算：
  - 关系型数据库→分布式关系型和非关系型数据存储计算→ 分布内存分析计算/硬件计算技术
- 大数据硬件：小型机→通用服务器硬件(X86)

技术发展方向：基于X86通用服务器，提供非结构化数据/结构化数据存储、融合分析能力，完成数据生命周期一体化管理，支撑业务准确决策、提升运作效率和业务创新能力。

目录

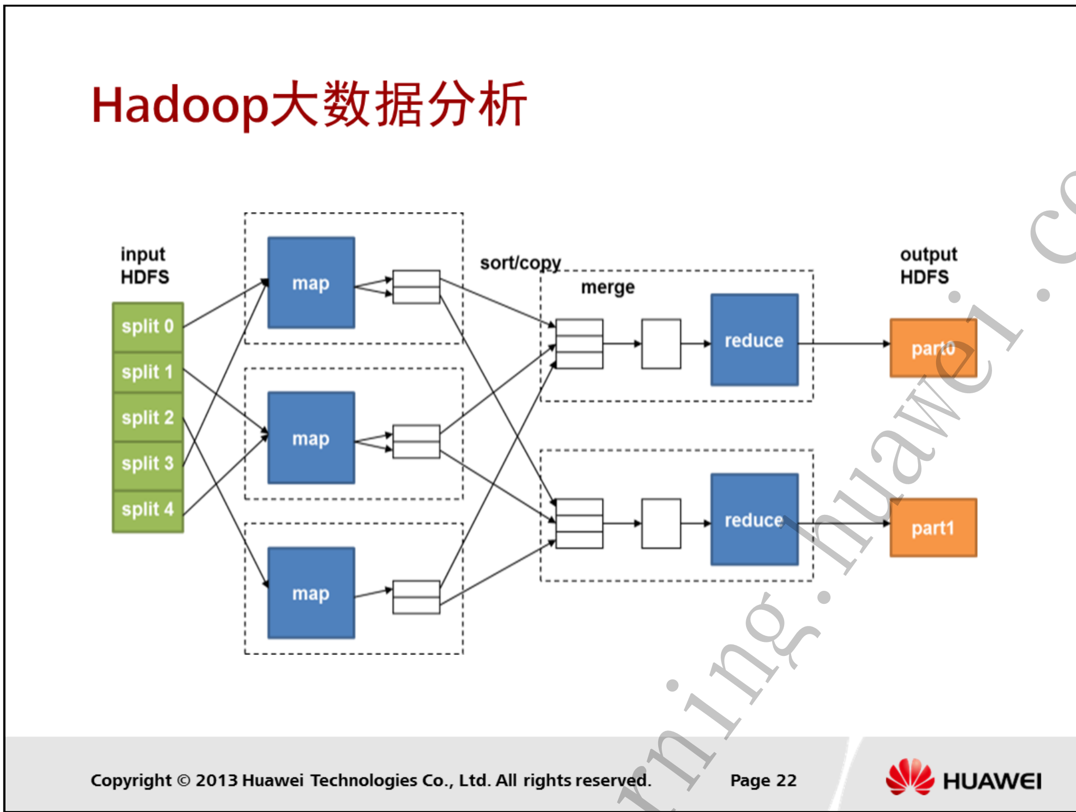


大数据存储解决方案
分布式存储
Hadoop FS API, NFS/CIFS,Posix,ISCSI,S3
Scaleout Block
Scaleout Filesystem
Scaleout DB
Scaleout Object
Scaleout backup
Cloud OS
通用服务器
网络（交换机，路由器）
存储节点
OceanStor 9000
数据全生命周期管理
Scale-out 文件系统
计算
集群
虚拟化
存储
网络
10GE
SAS
IB
数据库
分布式数据库
非关系型数据库
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 20
HUAWEI



针对以非结构化数据为主的互联网大数据处理技术比较有代表性的是Hadoop。Hadoop是一种开源的对大规模数据进行分布式处理的技术框架，在处理大数据中非结构化数据上有着性能和成本方面的优势。

Hadoop主要分为三个部分，分别是：HDFS（Hadoop Distributed File System）分布式文件系统、Hbase（Hadoop Database）非关系型数据库和MapReduce分布式并行处理架构。



Apache Hadoop 是一个用java语言实现的软件框架，在由大量计算机组成的集群中运行海量数据的分布式计算，它可以让应用程序支持上千个节点和PB级别的数据。Hadoop是项目的总称，主要是由分布式存储（HDFS）、分布式计算（MapReduce）等组成 。

HDFS是一个高度容错性的分布式文件系统，能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。

- NameNode：可以看作是分布式文件系统中的管理者，存储文件系统的meta-data，主要负责管理文件系统的命名空间，集群配置信息，存储块的复制
- DataNode：是文件存储的基本单元。它存储文件块在本地文件系统中，保存了文件块的meta-data，同时周期性的发送所有存在的文件块的报告给NameNode。
- Client：就是需要获取分布式文件系统文件的应用程序。

Map（映射）和Reduce（化简），采用分而治之思想，先把任务分发到集群多个节点上，并行计算，然后再把计算结果合并，从而得到最终计算结果。多节点计算，所涉及的任务调度、负载均衡、容错处理等，都由MapReduce框架完成，不需要编程人员关心这些内容。

- TaskTracker：Map/Reduce任务执行
- JobTracker：任务调度，监控

总结

- 大数据产生与特点
- 大数据对IT的需求和促进
- 大数据解决方案

思考题

- 大数据的特点有哪几个？与结构化和非结构化数据有什么关系？
- 大数据对IT或存储的需求有哪几个方面？
- 大数据解决方案的最终价值有哪些？

Thank you

www.huawei.com

HC1209202 大数据存储技术

HC1209202

第二章 大数据存储技术

www.huawei.com

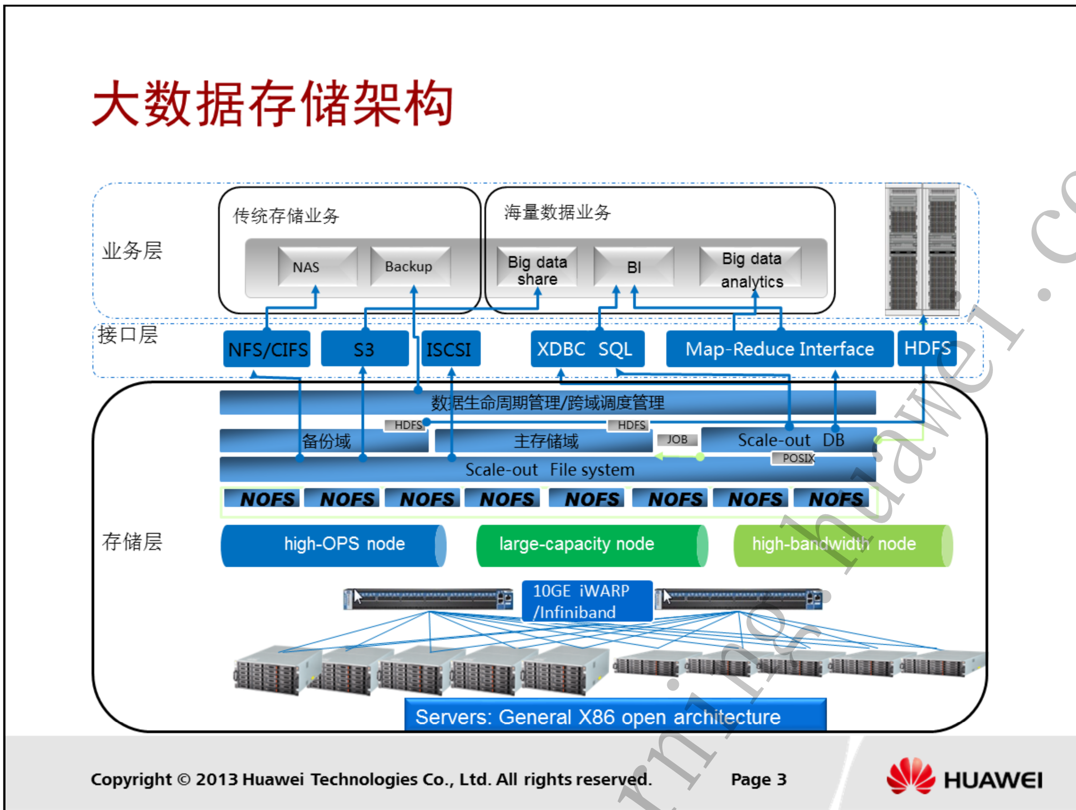
HUAWEI

目标

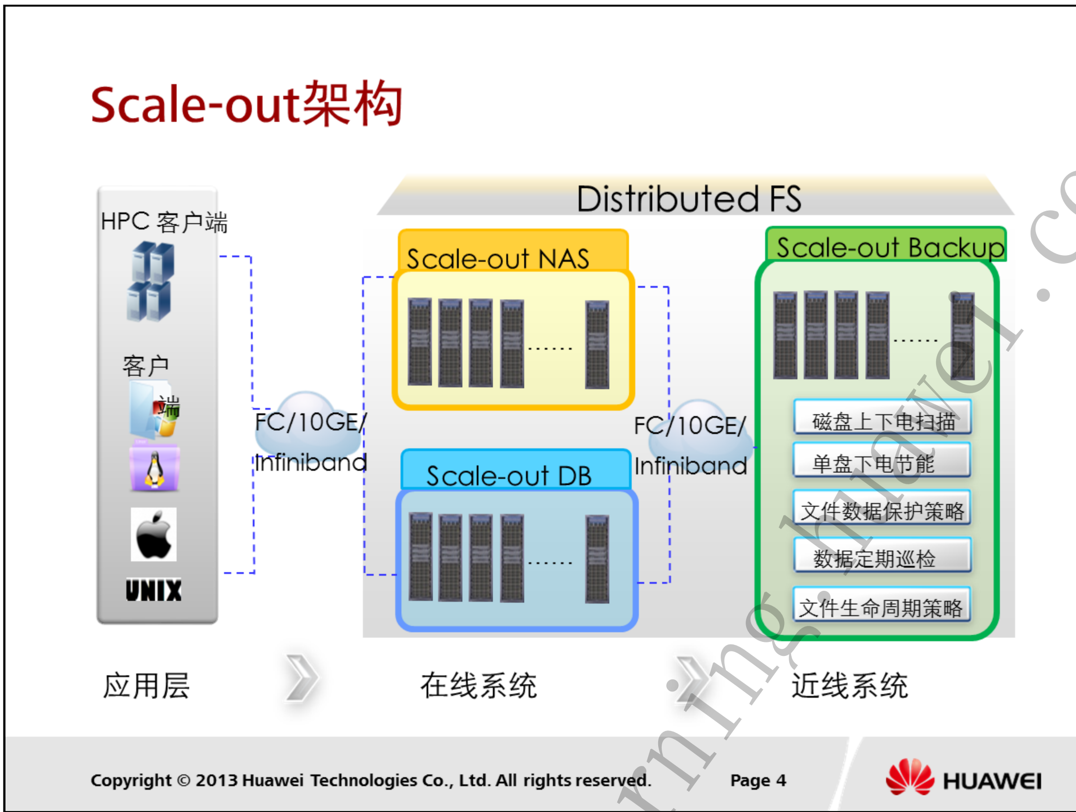
- 学完本课程后，您将能够:
  - 描述大数据存储架构
  - 理解大数据存储技术
  - 了解大数据分析技术
  - 了解数据安全技术
  - 理解集群NAS技术
  - 理解文件系统分级存储
  - 理解文件系统快照、镜像、复制

目录

  HUAWEI



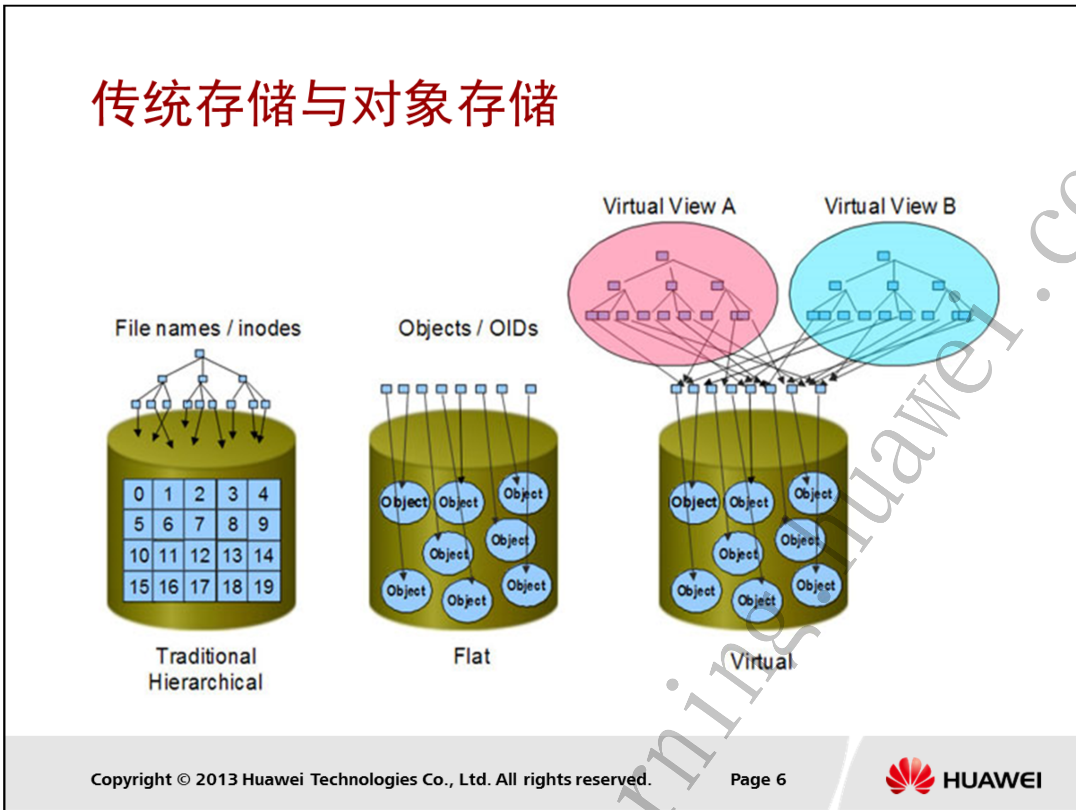
- 华为大数据存储解决方案OceanStor 9000架构特点：
  - 融合Scale-out NAS、Scale-out Database和Scale-out Backup，实现多位一体，面向大数据存储的集群存储系统。
  - 在一个系统内实现了分布式存储、分布式备份以及分布式数据分析的一体化全生命周期管理，在数据统一调度模块的调度下，数据在多域间有效流动。
  - 采取分布式架构，系统在初始时可以使用较小的配置，降低CAPEX(Capital Expenditure 基本建设费用支出)开销，随着业务量的增加，客户可以方便扩容，以实现应需而变。



- Scale-out是横向扩展的一种架构，这种架构主要应对数据的快速增长和复杂变化带来的需求和挑战。有关Sacle-out架构中的特征主要有：分布处理、在线扩展、数据共享、集群化。
- 分布式存储，分布式数据库都具备横向扩展能力。

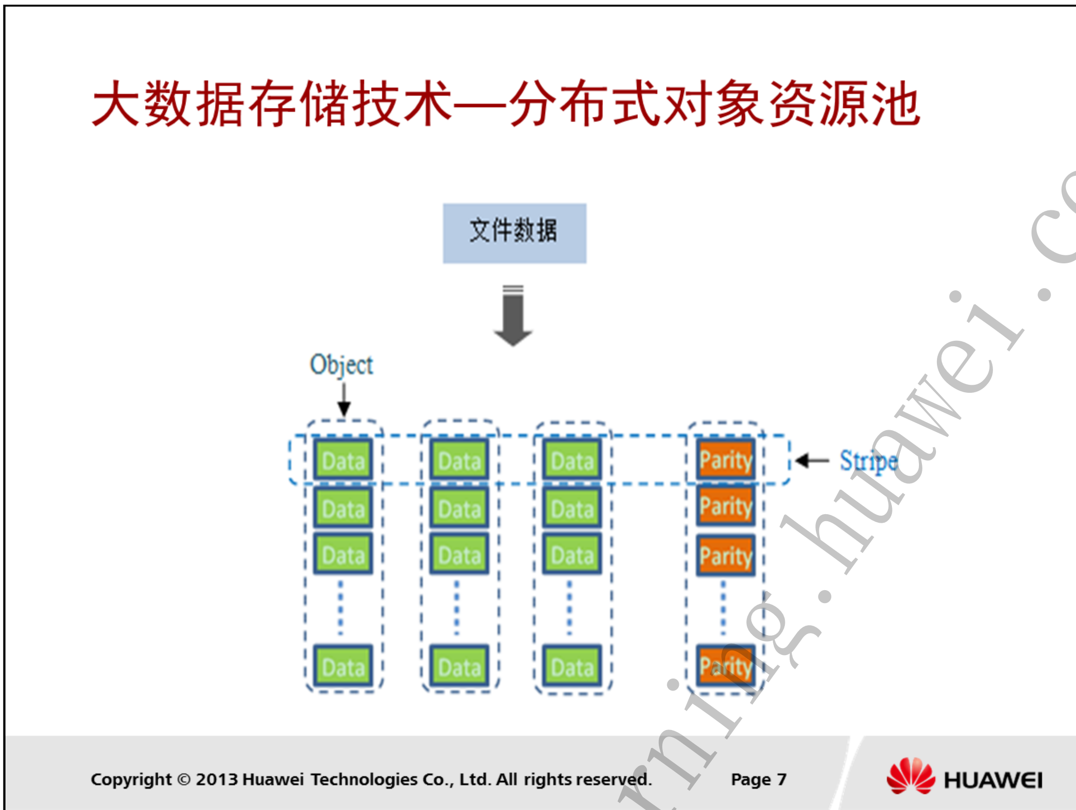
目录

1. 大数据存储架构
2. **大数据存储技术**
3. 大数据分析技术
4. 数据安全技术
5. NAS基础技术
6. 集群NAS技术
7. 集群NAS卷管理技术



传统数据存储方式是将磁盘映射成分块，文件系统管理分块，并使用元数据来标记分块使用情况。对象存储访问不关注分块，看到的是一个个的对象，上层使用对象名来访问。这种对象系统通常是扁平化的，区别于文件系统。

虚拟化借鉴了两种优势，一方面把底层以对象方式管理，简化了管理方式和元数据开销；另一方面，上层在虚拟出的对象基础上仍然构建成文件系统树形结构。

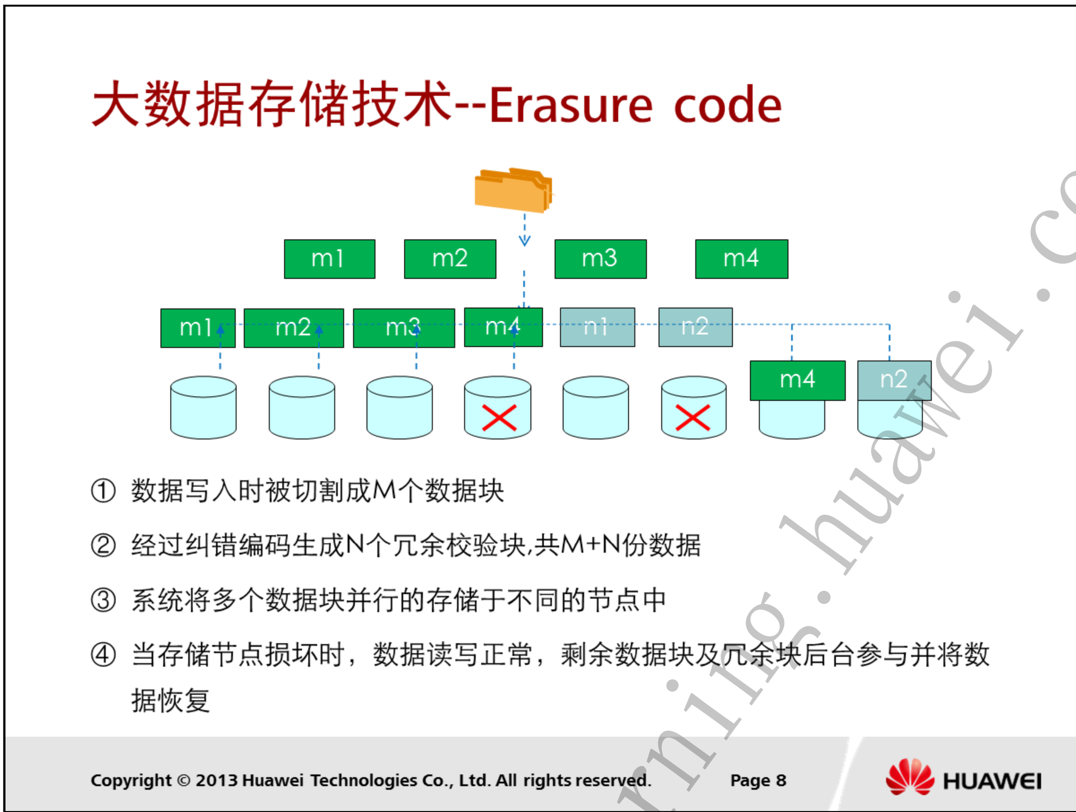


对象是对象存储的基本单元，每个Object是数据和数据属性集的综合体。数据属性可以根据应用的需求进行设置，包括数据分布、服务质量等。在传统的存储中，块设备要记录每个存储数据块在设备上的位置。Object维护自己的属性，从而简化了存储系统的管理任务，增加了灵活性。

OceanStor 9000的分布式文件系统，是以底层的集群对象存储系统为基础的。OceanStor 9000 集群对象存储系统是一种基于对象的存储系统，具有高并发、高吞吐量等优点。

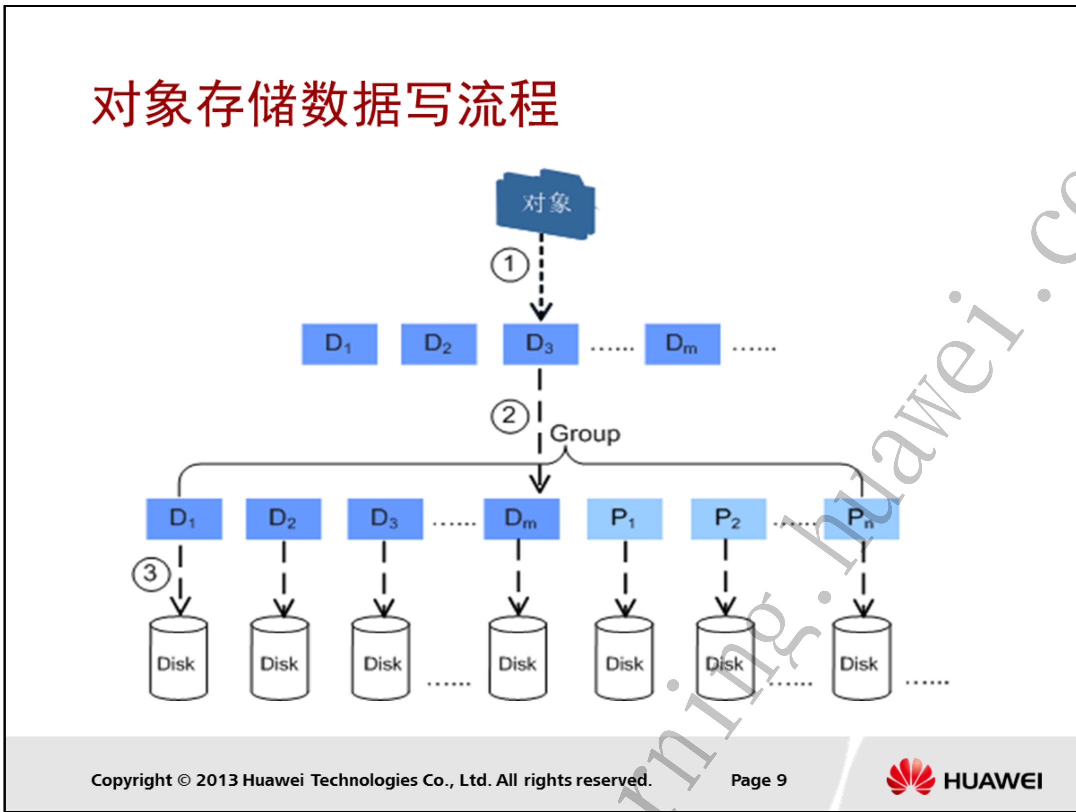
OceanStor 9000分布式文件系统的数据和元数据，经过数据条带化过程，生成的分片和条带，按列组合成对象，最终以对象的方式存储到集群对象存储系统中。以一个3+1方式保护的文件数据举例，如图中所示。

OceanStor 9000集群对象存储系统可以进行自管理。它不间断地监视着系统内硬盘的使用状况，当检测到有负载不均衡发生时，它会自动发起对象迁移，使系统重新达到负载均衡。当有硬盘或者节点损坏时，集群对象存储系统会自动发现故障，并自动发起对象级别的数据恢复。这些受损的对象会选择不同的节点和硬盘作为恢复目标，并发地执行恢复过程。这样，OceanStor 9000系统可以达到非常高的数据恢复速度，提高了数据的安全性。



Erasure code冗余技术是RAID的超集，能够支持比传统RAID算法更高的可靠性和更灵活的冗余策略。Erasure Code对文件进行原始分片（M份），将对象分割成固定长度的数据块，每M个连续数据块通过Erasure Code算法计算出N个校验块（N<M），这M个数据块和N个校验块组成一个数据块组并保存在不同的硬盘上。在Erasure Code存储模式下，系统最多支持N块硬盘失效，（M+N）/M=1.x即为保存的数据副本系数。

OceanStor 9000系统可以对目录/文件配置不同的数据保护方式。不同的数据保护方式，是通过不同的数据条带化方式实现的。使用Erasure Code方式时，写入OceanStor 9000系统的数据，会按照128KB为一个分片，切分为多个源数据分片，然后对每M个源数据分片，计算得到N个冗余数据分片，最终每M+N个分片组成一个条带，写入到系统中。当系统出现故障，丢失了其中的某些分片时，只要一个条带中丢失的分片数目不超过N，仍可进行正常的数据读写。通过数据恢复算法，丢失的分片可从剩余分片中计算得到。在这种方式下，空间的利用率约为 M/(M+N)，数据的可靠性由N值的大小决定，N越大可靠性越高。使用镜像方式时，写入OceanStor 9000的数据也将按照128KB进行切片，不同的是，系统对每个源数据分片，都会按照复制的方式存储为多份，只要其中任意一份不损坏，就可进行正常地读写。执行数据恢复时，丢失的镜像数据直接从完好的数据中复制即可。同样，系统可以配置不同的镜像数目，以得到不同的空间利用率和数据可靠性。另外，OceanStor 9000的分布式文件系统的元数据，采用的就是镜像的存储方式，并且对于配置了M+N保护方式的文件来说，其元数据的镜像数自动设定为N+2份，这使得OceanStor 9000分布式存储系统的元数据的可靠性高于其数据。

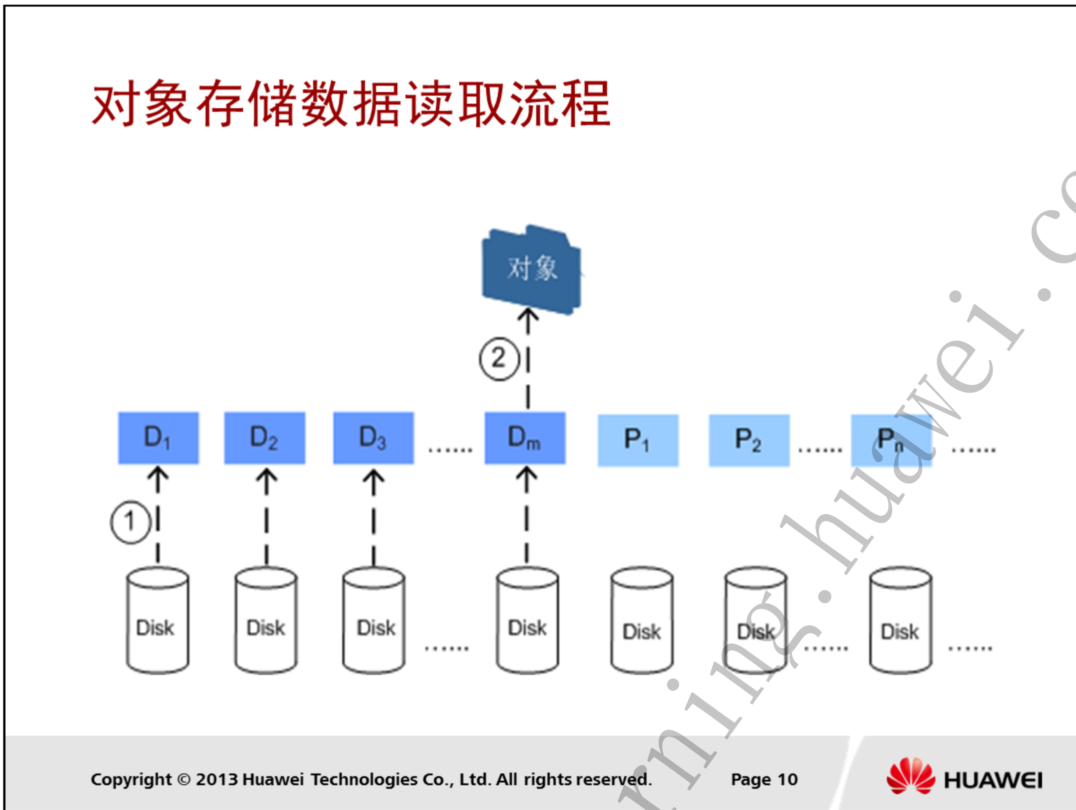


- 数据写的过程：

① 用户上传的对象数据被切分成若干固定长度的数据块。

② 每M个数据块通过Erasure Code算法计算出N个冗余校验块（N<M）。这M个数据块和N个校验块组成一个数据块组，图中的数据块D1～Dm和校验块P1～Pn组成一个数据块组。

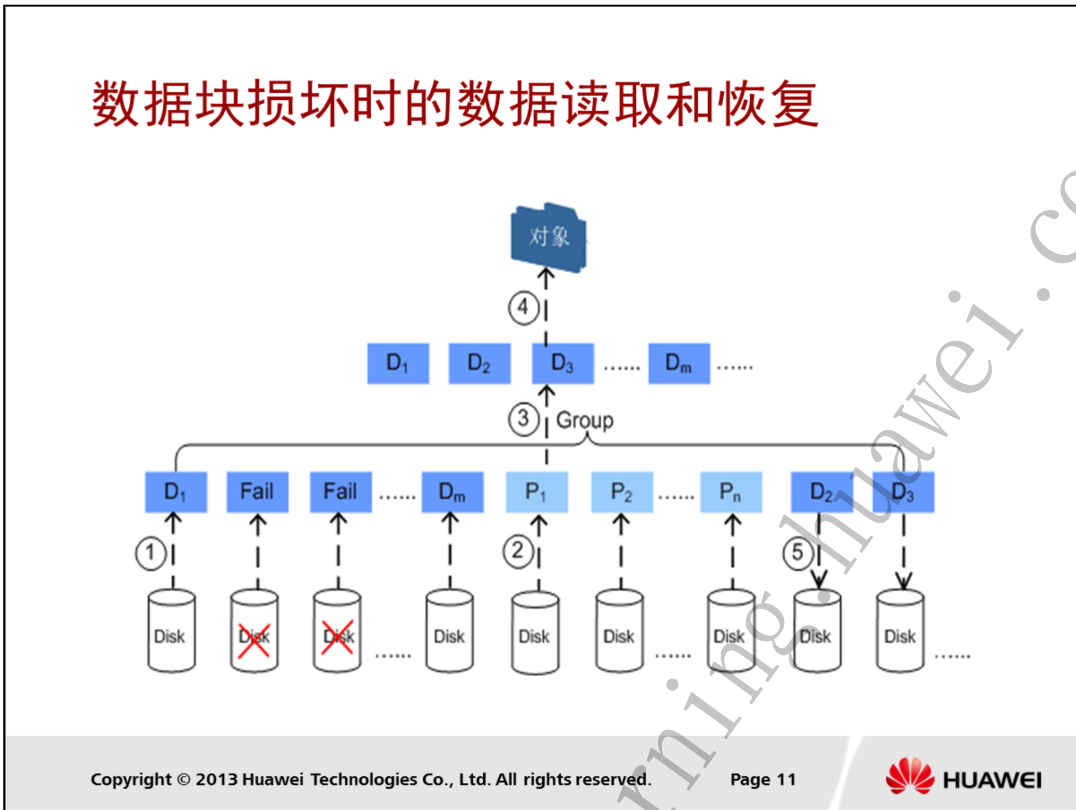
③ 数据块组中的数据块和校验块随机保存到不同存储节点的不同硬盘上。



- 数据读取过程：

① 系统读取保存于不同硬盘上的数据块。

② 系统将读取出来的所有数据块合并成完整的对象数据返回给用户。



- 数据恢复过程：

① 系统读取保存在不同硬盘上的M个数据块。

② 发现有数据块损坏，读取校验块。

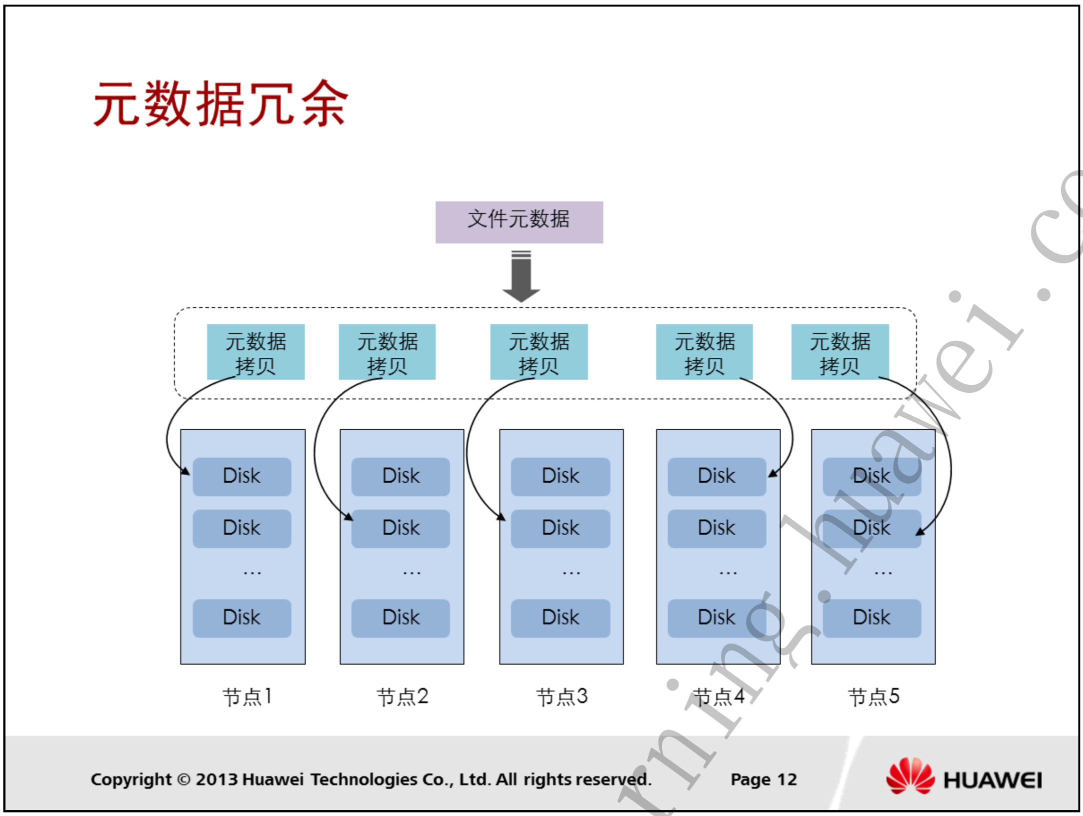
③ 根据读取出来的数据块和校验块，通过Erasure Code算法恢复损坏的数据块，然后将完整的M个数据块返回给系统。

④ 系统将读取出来的所有数据块合并成完整的对象数据返回给用户。

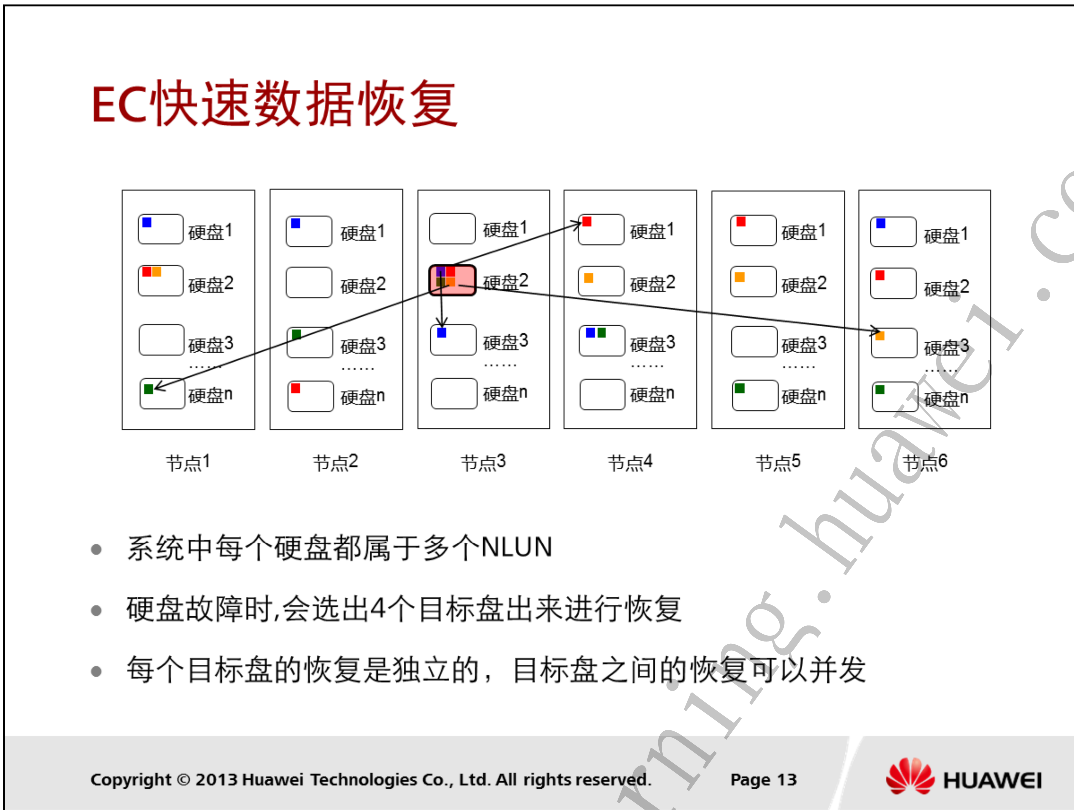
⑤ 系统将恢复出来的数据块存入其他正常工作的硬盘上。

如果单个数据块组中损坏的数据块和校验块总和大于N，系统会以更细粒度的数据块重新读取数据（例如：64kB，正常情况下，系统都是以1MB的数据块粒度读取数据），尝试恢复每个已损坏数据块中可读取的数据。通过这种数据部分修复的方式，系统可以最大程度地减少损坏的数据块和校验块数量。

注意：如果大于N个数据块和校验块损坏且系统通过部分修复方式无法使损坏的数据块和校验块数量小于或等于N，丢失的数据将无法恢复。



元数据和数据一样都是使用节点间冗余方式，通过镜像实现，每一份都是一个独立完整的拷贝。



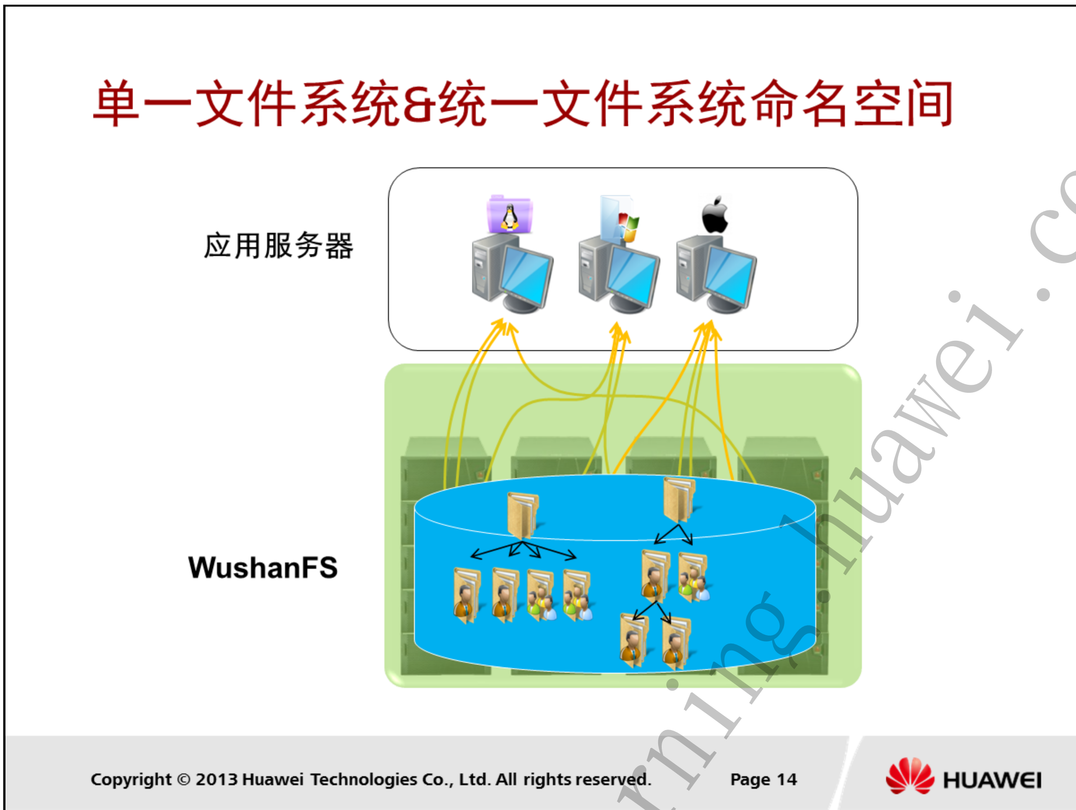
以图中节点3硬盘2为例，同时属于4个NLUN（四个不同色块），当这个盘故障时,会选出4个目标盘出来进行恢复，数据会分别在4个目标盘上被恢复出来。由于每个目标盘的恢复是独立的，目标盘之间的恢复可以并发，所以恢复速度较高(阵列中类似的故障下目标盘只有一个)。在实际系统中，一个盘所属的的NLUN数将超过12个，可达到每小时1TB的恢复能力。

同时，可以配置任意时间段内恢复的优先级，优先级分为高、中、低三档，默认优先级是中。优先级体现了数据恢复对系统性能的影响，优先级越高对系统性能影响越大。

- 高优先级：影响不超过50%
- 中优先级：影响不超过30%
- 低优先级：影响不超过10%

当发生节点或硬盘故障时，故障节点所提供的服务会启动迁移至集群中的其他节点，保证业务正常运行。当故障节点或硬盘恢复时，OceanStor 9000可通过节点间数据重构的方式快速恢复所丢失的数据，保证整系统的数据高可靠性。包括以下几种情况：

- 若硬盘故障后在10分钟内恢复
- 系统自动启动硬盘数据恢复，且仅恢复故障期间变化的数据
- 若硬盘故障后超过10分钟恢复
- 系统自动启动硬盘数据恢复，但需恢复故障前已存储数据和故障期间变化的数据
- 节点故障恢复后，系统自动启动节点数据恢复，且仅恢复故障期间变化的数据
- 过网管手工执行全量数据恢复



OceanStor 9000对外提供统一的文件系统命名空间。用户存储文件、设置共享、管理权限的操作对象均位于该命名空间内。统一的命名空间可以使用和管理当前系统的全部可用容量。

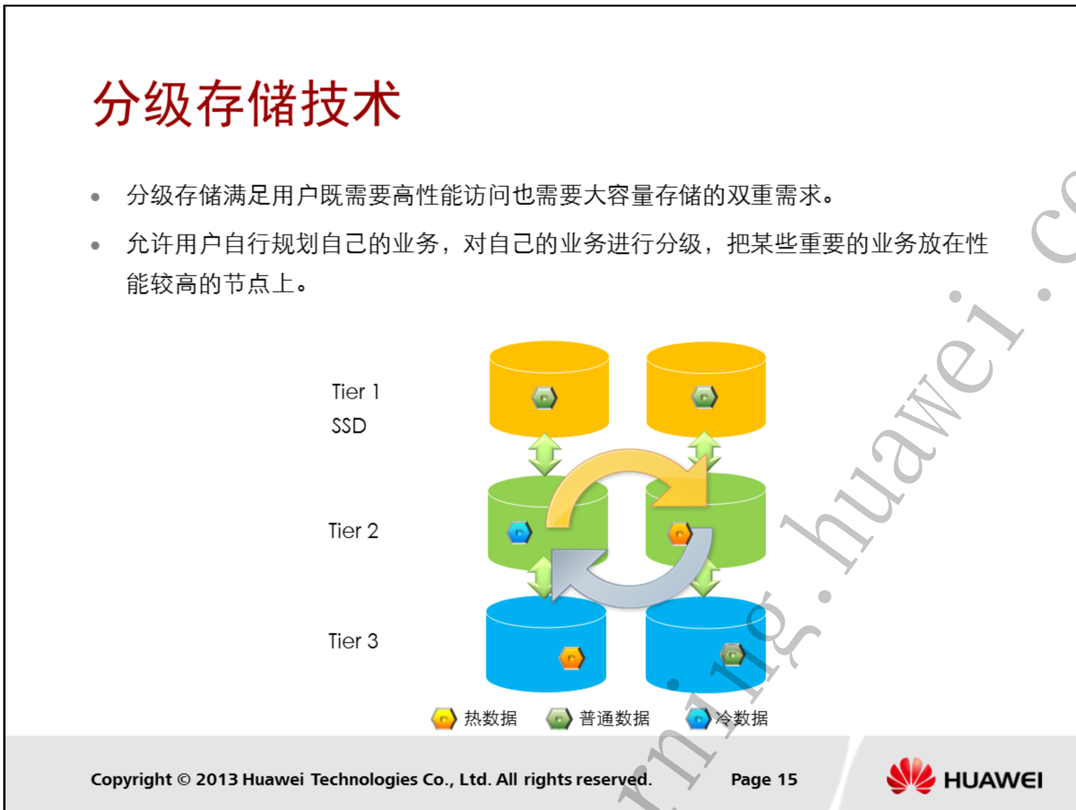
命名空间在OceanStor 9000启动时自动创建，其名称与系统命名保持一致。各共享目录均为命名空间的子目录。同时命名空间作为整个系统的根目录，也可以被共享和访问。

统一的命名空间为用户提供统一的存储资源池，实现对存储空间的集中管理和分配。当进行扩容时，新增容量可直接纳入资源池使用，避免维护多个文件系统。

Wushan分布式文件系统对外提供CIFS/NFS访问接口，即最终用户以文件名来操作文件，那么文件命名的有效范围就称之为命名空间。

通常一个本地文件系统，如ext3文件系统，一个命名空间只能与一台计算机或文件系统相关联，本地文件系统将元数据和数据都存储在本地磁盘，访问文件数据时文件系统首先从本地磁盘读取元数据信息，然后通过元数据信息得知文件数据在本地磁盘存储的布局和位置，再去读写本地磁盘完成访问文件操作。

Wushan分布式文件系采用全对称结构，将整个系统的所有节点的存储容量组成一个大的虚拟存储池，元数据和数据保存在每个节点上，每个节点都是元数据服务器同时也是数据服务器，访问文件数据时Wushan文件系统首先定位到文件所归属的元数据服务器，然后通过元数据服务器获取文件的数据分布，即获取文件分布在哪些节点上以及在节点上的具体位置，再去访问去这些节点完成数据读写操作。

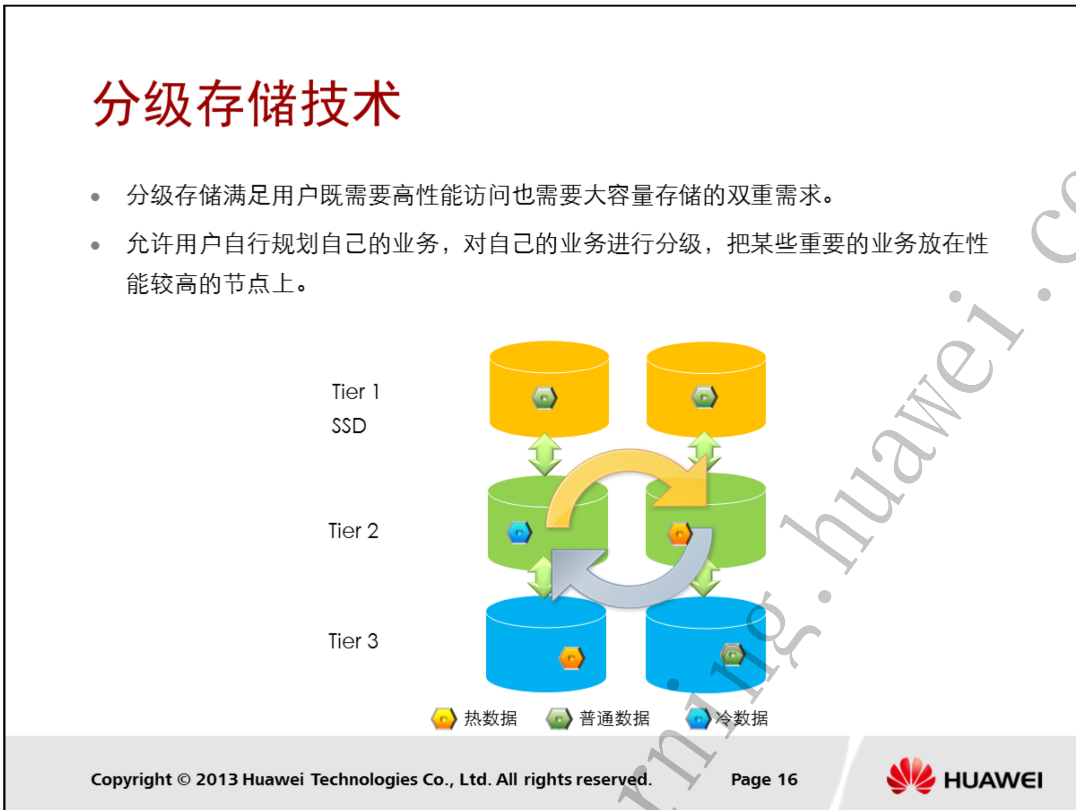


- 动态分级存储的特点：
  - 支持不同类型节点间分级存储
  - 基于自定义的分级策略
  - 支持文件大小、路径、时间属性、IO热度等策略
  - 无缝地适应业务变更，而不会影响应用程序
  - 将多个应用程序整合到单个存储系统中，消除了数据移动且简化了管理

OceanStor 9000动态分级存储允许将同一个文件系统内的物理节点划分成不同的节点池。节点池是指具有相同特征(物理类型/访问性能)的节点的集合。文件池是符合某些特征的文件的集合，而文件池策略则是定义文件池如何在节点池中存放的规则。

动态分级存储允许用户基于文件池策略定义工作流中数据的价值。将高价值(重要)的文件放置在高可用性、高性能（相应地，成本也非常高）的存储设备上；低价值(不太重要)的文件放置在成本较低的、性能和可用性规格较低的设备上。动态分级存储是根据文件池策略触发，自动进行数据的迁移，对业务访问透明，不会对业务造成影响。系统支持根据物理节点类型自动生成节点池，包括高性能节点池，高带宽节点池，大容量节点池 。在部署和扩容时系统会按照节点所属的物理类型将节点加入到对应的节点池。支持手工增删节点池，在部署或扩容时将节点加入到指定的节点池中。 支持手工增删节点池，在部署或扩容时将节点加入到指定的节点池中。

在系统运行过程中允许对已经加入节点池的节点进行手工调整，将其从一个节点池移到另外一个节点池中。具体做法为将节点从现有节点池中删除，再将该节点加入到另一节点池中。需要注意的是当前节点池中的节点个数必须能满足已经配置的保护级别N+M个，否则会因为会降低当前数据的保护级别而无法删除，在此种情况下只能删除整个节点池。

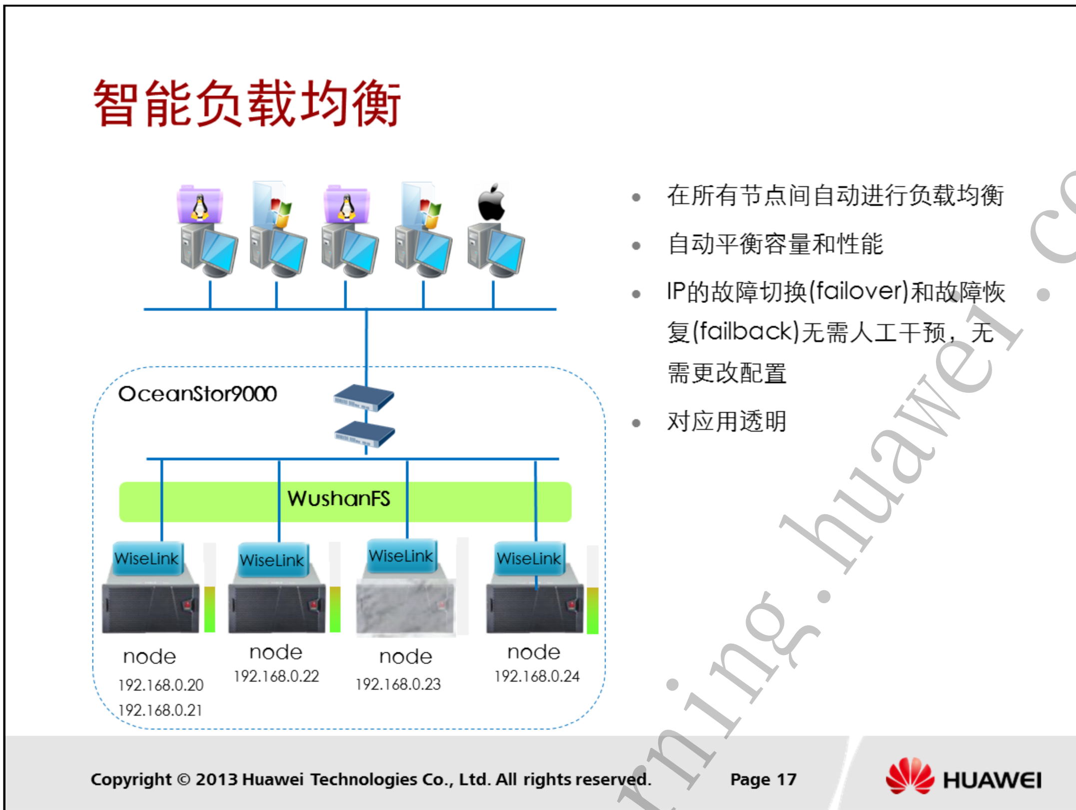


系统最多可以配置64个节点池，每个节点池至少包括3个物理节点，否则该节点池的状态为“不可用”。　建议节点池中的节点为同一物理类型的节点，否则该节点池的性能将会收到影响。系统会统计节点池磁盘空间的使用情况，如果空闲的磁盘空间超过系统设置的低水位，则不允许新的文件存放/迁移到该节点池。

系统同时也会产生告警，提醒用户进行扩容或者修改文件池策略。 系统支持基于文件属性以及I/O热度定义文件池策略，包括文件存放路径，文件名，文件大小，创建时间，修改时间，访问时间，UID/GID，I/O热度，I/O访问次数。文件池策略包括文件特征定义以及存放该文件池的节点池ID/名称。

文件池策略分为3种：create，relocate，delete。

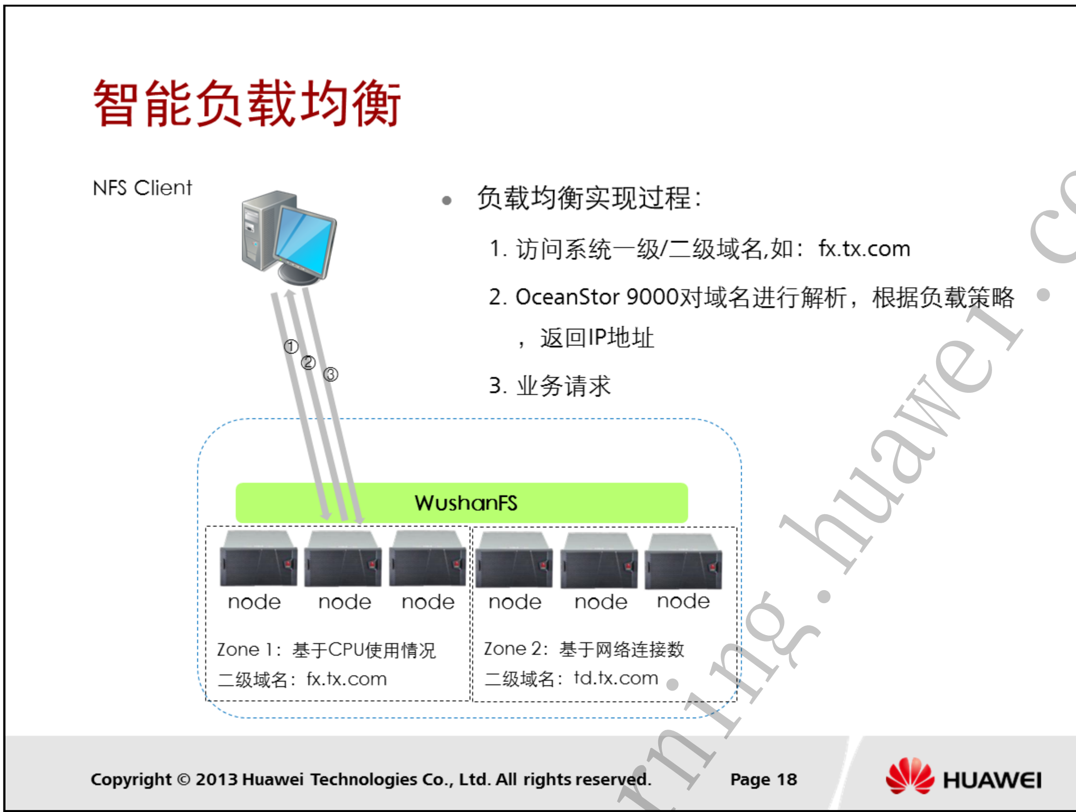
- Create策略用于创建文件时确定文件的初始放置位置，该类策略条件包括：存放路径，文件名，UID/GID。可以配置默认文件池策略，该策略无任何条件，只有对应的节点池。缺省策略用于确定无法匹配其他create文件池策略的文件存放位置。
- Relocate策略用于确定文件迁移的目标位置。该类策略条件包括：文件存放路径，文件名，文件大小，创建时间，修改时间，访问时间，UID/GID，I/O热度，I/O访问次数。这些条件之间是“与”的关系，不输入表示不匹配该属性，只有同时满足这些输入的条件的文件才会迁移到对应的节点池中。
- Delete策略用于确定何时将文件从系统中删除。需要注意的是该策略使能后文件将不能恢复，需要谨慎使用配置该策略。该类策略的条件为：存放路径，创建时间，修改时间，访问时间，UID/GID。



负载均衡（又称为负载分担），英文名称为Load Balance，其意思就是将负载（工作任务）进行平衡、分摊到多个操作单元上进行执行，例如Web服务器、FTP服务器、企业关键应用服务器和其它关键任务服务器等，从而共同完成工作任务。负载均衡建立在现有网络结构之上，它提供了一种廉价有效透明的方法扩展网络设备和服务器的带宽、增加吞吐量、加强网络数据处理能力、提高网络的灵活性和可用性。

常见的负载均衡技术有如下几种：

- DNS负载均衡：最早的负载均衡技术是通过DNS来实现的，在DNS中为多个地址配置同一个名字，因而查询这个名字的客户机将得到其中一个地址，从而使得不同的客户访问不同的服务器，达到负载均衡的目的。
- 代理服务器负载均衡：使用代理服务器，可以将请求转发给内部的服务器，使用这种加速模式显然可以提升静态网页的访问速度。然而，也可以考虑这样一种技术，使用代理服务器将请求均匀转发给多台服务器，从而达到负载均衡的目的。
- 地址转换网关负载均衡：支持负载均衡的地址转换网关，可以将一个外部IP地址映射为多个内部IP地址，对每次TCP连接请求动态使用其中一个内部地址，达到负载均衡的目的。
- 协议内部支持负载均衡：除了以上三种负载均衡方式之外，有的协议内部支持与负载均衡相关的功能，例如HTTP协议中的重定向能力等，HTTP运行于TCP连接的最高层。



- 负载均衡特点:
  - 对外提供单一域名访问方式
  - 可划分分区，对不同分区设置二级域名，以及不同的负载策略
  - 负载均衡策略可以基于：CPU、内存、带宽、网络连接数、节点能力值

OceanStor 9000的负载均衡服务目前支持如下的负载均衡策略，可以供用户按照实际的环境进行配置。

- 轮询方式

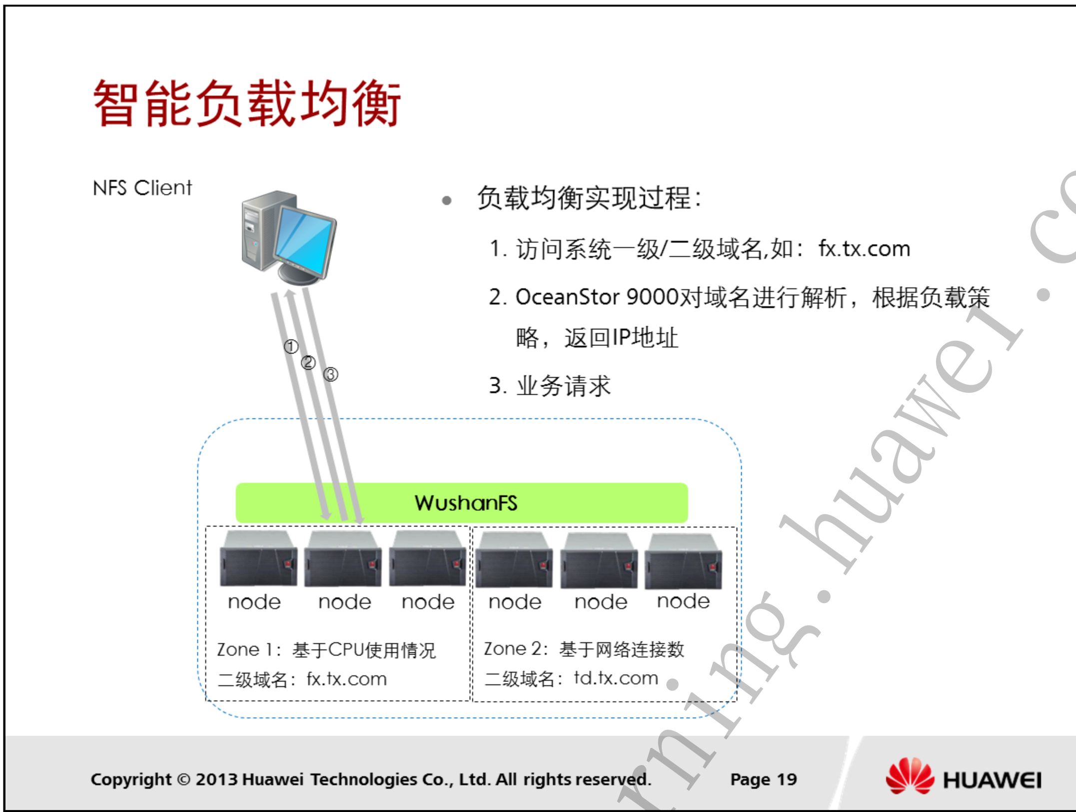
按顺序依次选择节点处理客户端连接请求。如图所示，以3节点环境为例。Tn代表DNS请求的时间点，那么轮询方式的负载均衡服务在T1时刻选择n1节点处理业务，T2时刻则选择n2节点处理业务，T3时刻选择n3节点处理业务。

- CPU使用率

选择CPU使用率最低的节点处理客户端连接请求。如图所示，以3节点环境为例。Tn代表DNS请求的时间点，那么按CPU使用率的负载均衡服务在T1时刻选择n1节点处理业务，T2时刻则选择n2节点处理业务，T3时刻选择n3节点处理业务。

- 节点连接数

选择NAS服务连接数量最低的节点处理客户端连接请求。如图所示，以3节点环境为例。Tn代表DNS请求的时间点，那么按节点连接数的负载均衡服务在T1时刻选择n1节点处理业务，T2时刻则选择n2节点处理业务，T3时刻选择n1节点处理业务。



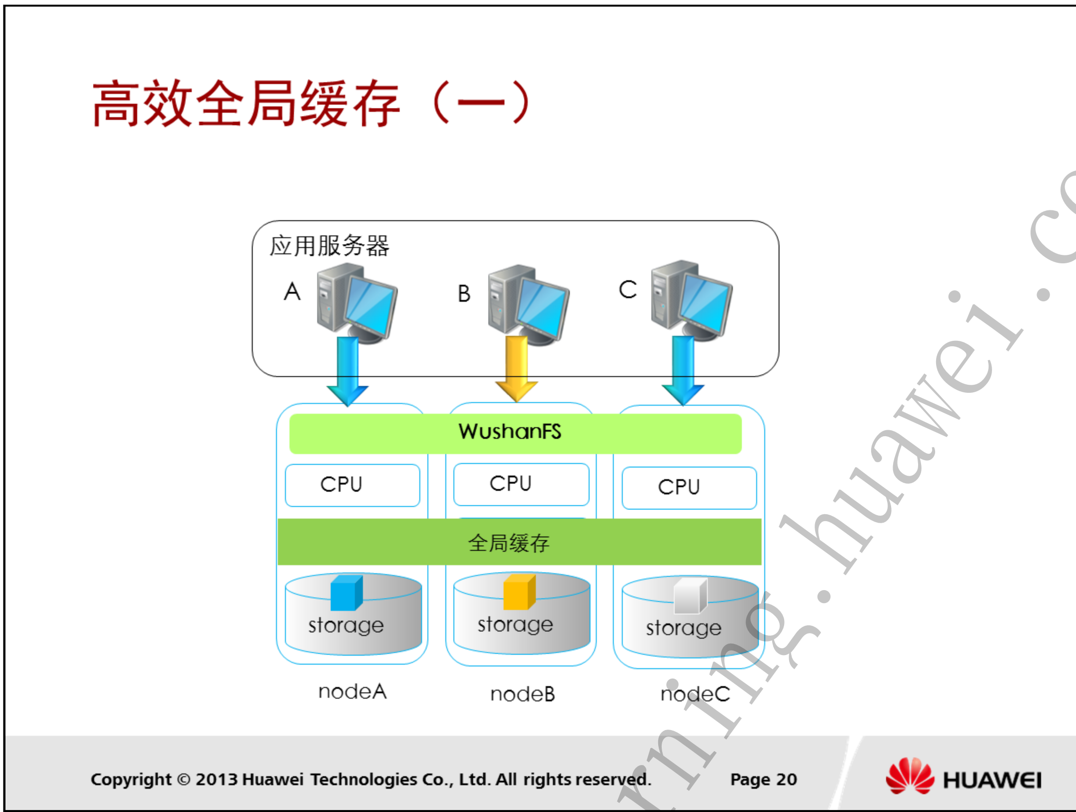
- 节点吞吐量

选择网络吞吐量最低的节点处理客户端连接请求。如图所示，以3节点环境为例。Tn代表DNS请求的时间点，那么按节点连接数的负载均衡服务在T1时刻选择n2节点处理业务，T2时刻则选择n1节点处理业务，T3时刻仍选择n1节点处理业务。

- 节点能力

负载均衡服务根据节点的CPU、带宽等硬件配置计算节点能力值，根据节点当前的运行状态值计算节点的负载状态。如果某节点负载过重则自动减少该节点的能力值。如果当前负载比较轻，节点能力也会得到相应的增加。

OceanStor 9000最大规格支持288个节点，负载均衡系统允许IT管理员，对OceanStor 9000所有节点进行分区管理。可以按照需要对节点进行分类，对于每个分区负载均衡系统支持配置独立的负载策略、独立的访问域名。常见场景，IT管理员可以划分高性能分区和高容量分区，并给各个分区分配指定能力的节点，配置独立的域名。客户按照需要使用不一样的域名访问不一样的分区。如下图所示，4个节点可以划分为2个分区，其中一个为高性能分区配置域名为highperformance.OceanStor 9000.com；另外一个高容量分区配置域名为highcapactiy.OceanStor 9000.com。客户可以按照需要使用不同的域名访问不同的分区。

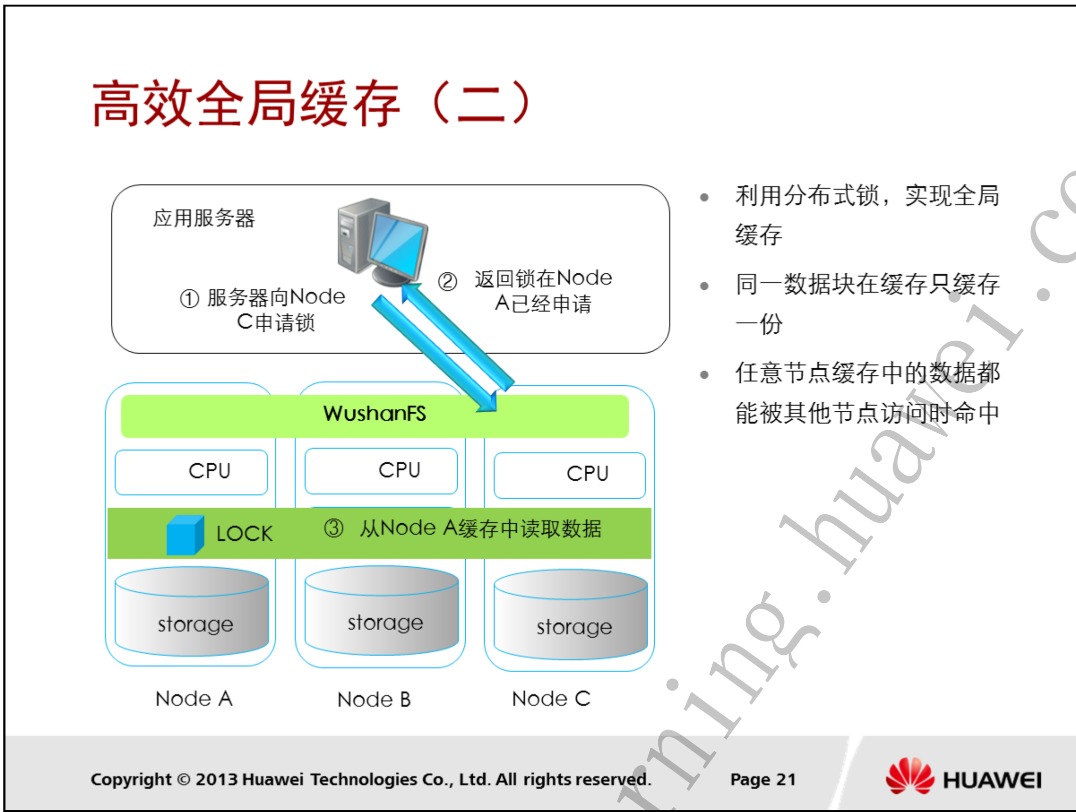


随着计算机技术、网络技术及人类生活信息化的发展，人们对存储系统的存储容量、存储性能的要求越来越高，存储系统由原来的计算机自带存储器发展到后来的存储阵列和网络存储（NAS）等独立存储存储系统，直到当前大型的高并发的分布式存储技术。

随着高清超高清视频监控、3D电影电视的出现和发展，视频在线编辑、高性能计算等应用对存储系统的访问性能提出了更具挑战性要求。

当前在SAN或NAS存储系统中，每个存储服务器上的缓存空间都是独立的，每个存储服务器只能访问缓存在本节点上的数据，造成不同的存储服务器上缓存相同的数据导致内存利用率低，无法充分发挥缓存的优势等问题。

WushanFS中的Global Cache使所有存储服务器上的内存空间在逻辑组成一个整体内存资源池，缓存在任何一个存储服务器上的数据，在后续其它任何存储服务器接收到访问该数据请求时都可以在全局缓存中命中，同时所有用户数据在整个集群系统中只缓存一份并且不缓存校验数据，可以有效发挥缓存对系统性能的影响提高系统整体的性能。



① 应用服务器向节点C端发出数据读请求时，首先应用服务器向巫山文件系统申请分条资源读锁

② 巫山文件系统会检查所读数据的缓存是否在全局缓存中以及缓存在哪个节点上，如图显示该文件分条资源在NodeA节点上的缓存中

③ 直接从NodeA节点上的全局缓存中获数据并返回（如果不在全局缓存中，则直接从各个节点上读取该分条数据的所有条带数据，然后构造出分条数据后再读取）



Wushan分布式文件系统支持节点动态扩展，而且节点扩展中业务不中断。Wushan分布式文件系统支持针对不同的应用需求提供不同的硬件节点类型，主要有支持高频度IO访问的高性能节点、满足带宽型应用的高带宽节点和支持海量需求的大容量节点。

由于采用了去中心化的技术，Wushan分布式文件系统不存在性能瓶颈，随着节点数的增加，存储容量和计算能力线性增加，最终给用户提供呈线性递增的带宽、并发数。

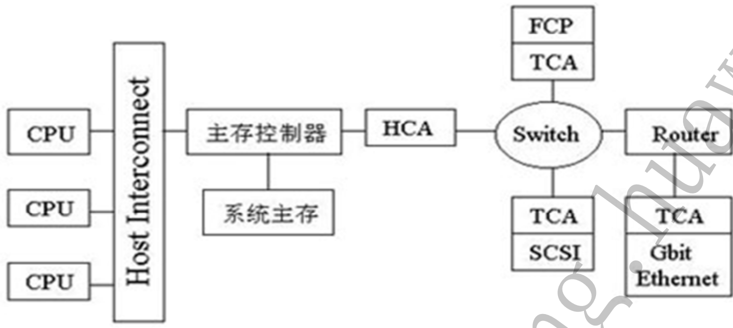
Wushan分布式文件系统提供了全局一致的缓存，缓存容量随着节点增加而线性增长，随着节点数目的增加，越来越多的热点数据可以被缓存命中，大大减少硬盘随机I/O，提高整系统性。



Infiniband协议的主要特点是高带宽（现有产品的带宽4xDDR 20Gbps,12x DDR 60Gbps, 4xSDR 10Gbps, 12xSDR 30Gbps、预计两年后问世的QDR技术将会达到4xQDR 40Gbps,12x QDR 120Gbps）、低时延（交换机延时140ns、应用程序延时3μs、一年后的新的网卡技术将使应用程序延时降低到1μs水平）、系统扩展性好（可轻松实现完全无拥塞的数万端设备的Infiniband网络）。

InfiniBand体系架构

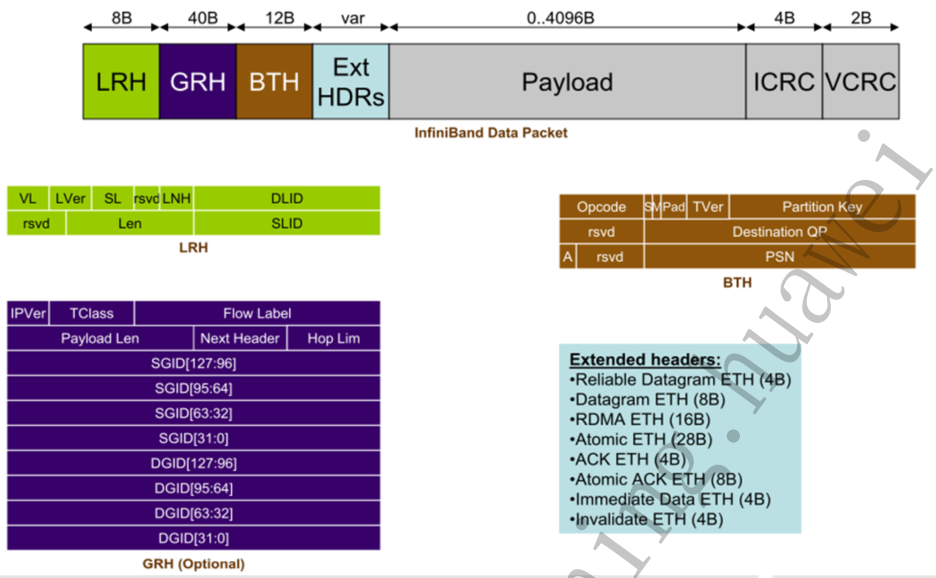
- InfiniBand标准定义了一套用于系统通信的多种设备，包括信道适配器、交换机和路由器。



 Page 24 HUAWEI

- 信道适配器用于同其它设备的连接，包括主机信道适配器（HCA）和目标信道适配器（TCA）
- 交换机是 InfiniBand结构中的基本组件
- 点到点的交换结构：解决了共享总线、容错性和可扩展性问题;具有物理层低功耗特点和箱外带宽连接能力

InfiniBand的帧结构

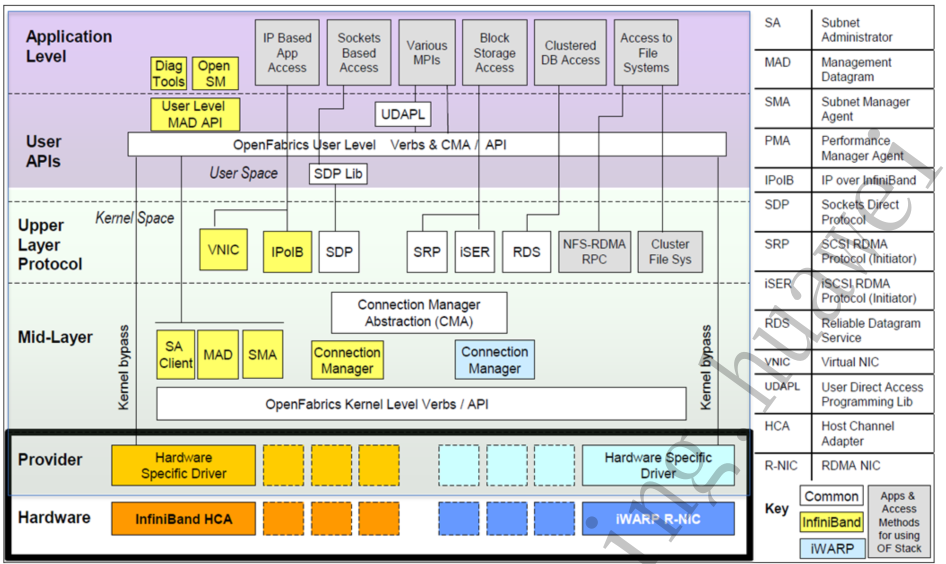


InfiniBand的特点

- 高速度
  - Infiniband第一代DDR技术，所支持的吞吐量为5、20或60Gbps，延迟低于1.3us。第二代QDR技术，带宽最高可达120Gbps，延迟低于100ns
- 远程直接内存存取功能
  - 该功能对于集群来说很适合，因为它可以通过一个虚拟的寻址方案，让服务器知道和使用其他服务器的部分内存，无需涉及操作系统的内核。
- 传输卸载
  - 远程直接内存存取能够帮助传输卸载，后者把数据包路由从OS转到芯片级，节省了处理器的处理负担

  HUAWEI

InfiniBand的协议栈



HUAWEI

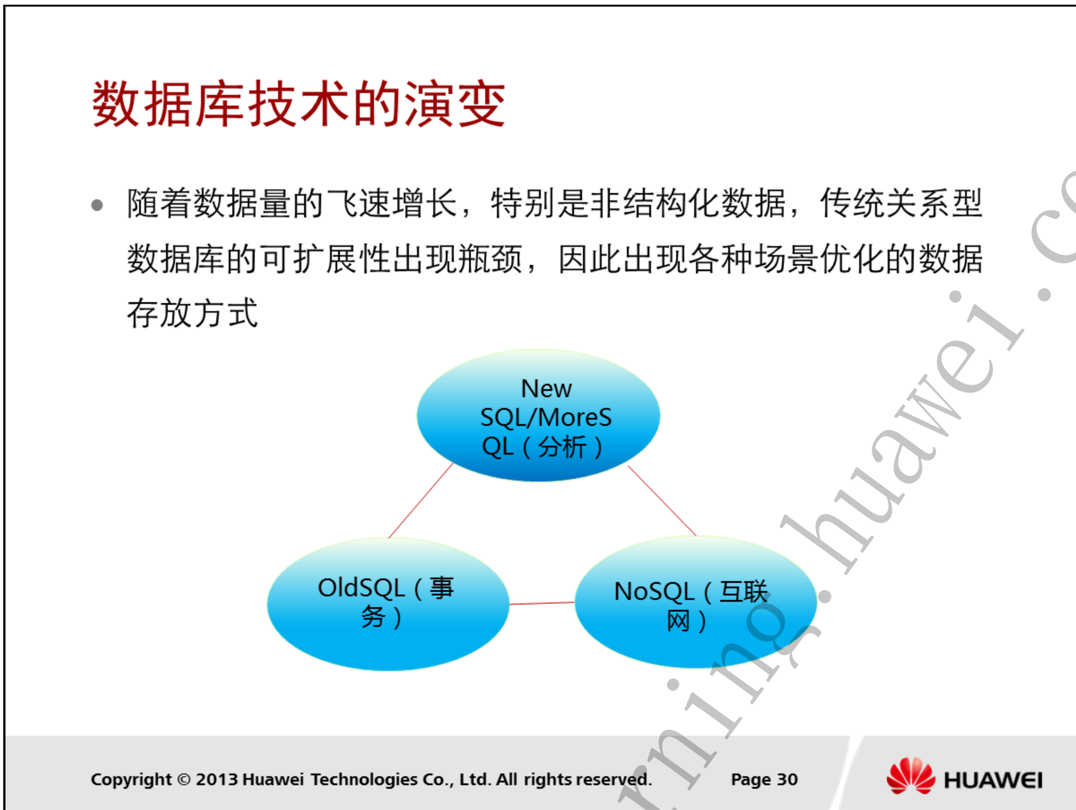
10GE以太网

- 定义
  - 10GE 以太网（10 Gigabit Ethernet，缩写为10GE，10GbE，10 GigE或10GE），即万兆以太网。
  - 一种以太网的传输标准，2002年通过，成为IEEE Std 802.3ae-2002。它规范了以10 Gbit/s 的速度来传输的以太网，它的速度是GE的十倍。
- 特点：
  - 10 Gigabit以太网，以全双工方式连接到网络交换器，它不支持半双工模式与CSMA/CD。
  - 10GE提供了高带宽低时延的访问，适合比较 媒资类、高算类、OLAP等。

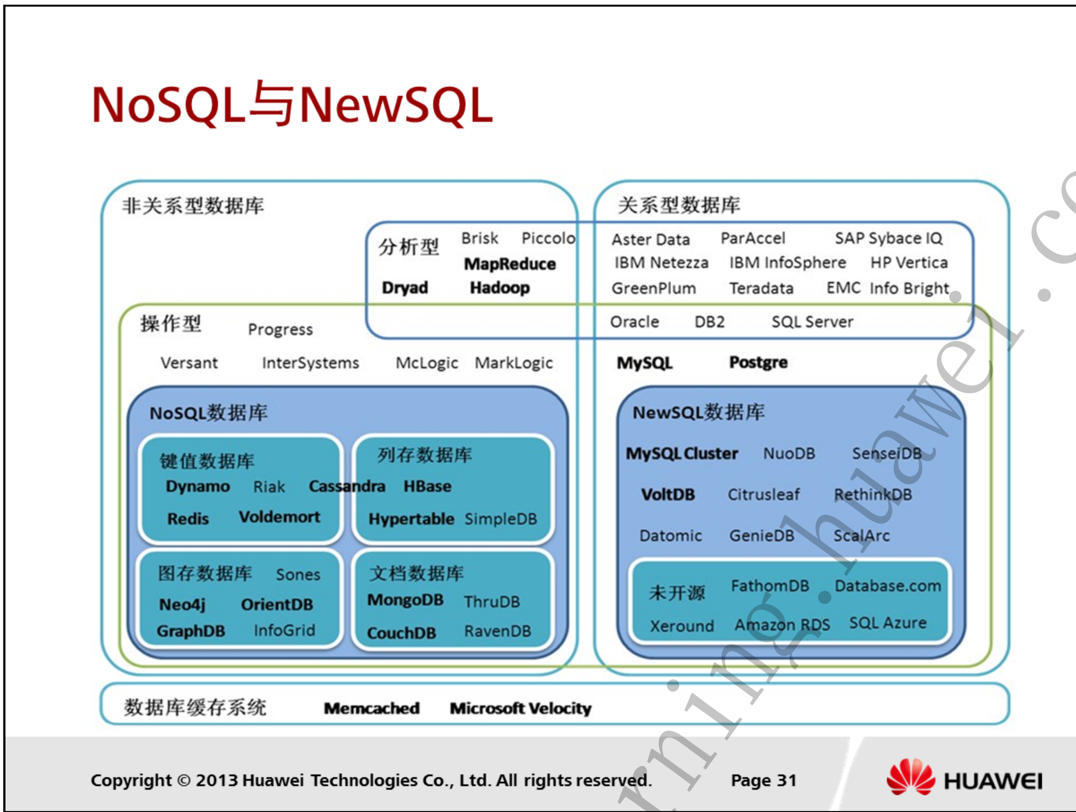
 Page 28 HUAWEI

- 10GE/FCOE/TOE对比：
  - 存储阵列主要采用3种板卡是10Gbit/s速率传输，分别是10GE、FCOE、TOE，其中又分为主机侧和阵列侧，主机侧为HBA卡，阵列侧是相应的扣卡，在此简单说明下各个板卡区别及使用。
  - 10GE：通过光纤信号进行传输，基于TCP/IP协议，上层是iSCSI协议，主机侧和阵列侧都有10GE接口卡，为双端口。
  - FCOE：通过FC光纤进行传输，但本质上使用的是以太网物理传输协议传输FC数据。FCOE相比于10GE来说节省了线缆、交换机，方便管理。主机侧和阵列侧都有FCOE接口卡，主机侧一般2端口/板，阵列侧4端口/板。
  - TOE：TOE相比于10GE，将TCP/IP协议处理放在了板卡上。相比于FCOE，可以选择走FC，也可以选择走ISCSI。阵列侧有TOE卡，4端口/板，主机侧无TOE卡。
  - 在主机侧和阵列侧连接的时候，可以选择不同类型的接口卡（FC/iSCSI/FCoE等），但是主机侧和存储侧需要保持通信协议一致。

目录

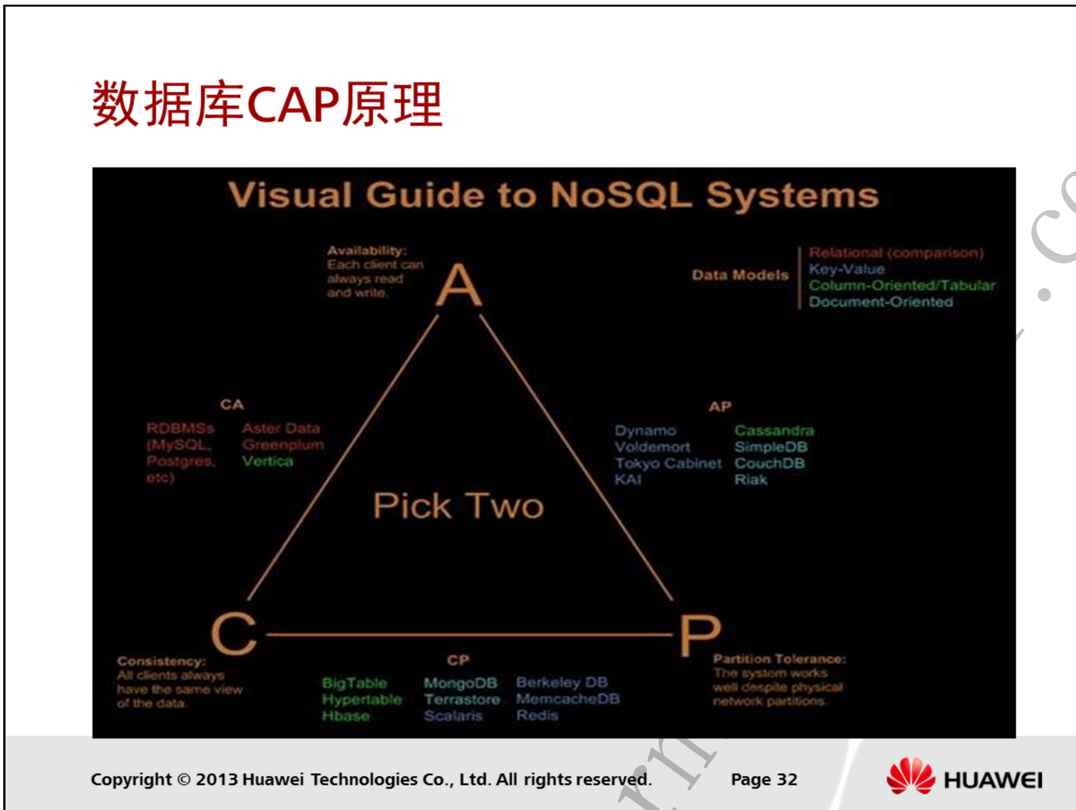


- NoSQL 的出现破除了关系型数据库一家独大的局面，代表为Cassandra、Riak、MongoDB、CouchDB、Hbase等
- MoreSQL,NewSQL出现，展现出比NoSQL更好的适应性
- 随着大内存和非易失型内存的出现，内存计算越来越受到重视
- 多种架构支持多种应用，OldSQL、NOSQL和NewSQL互为补充

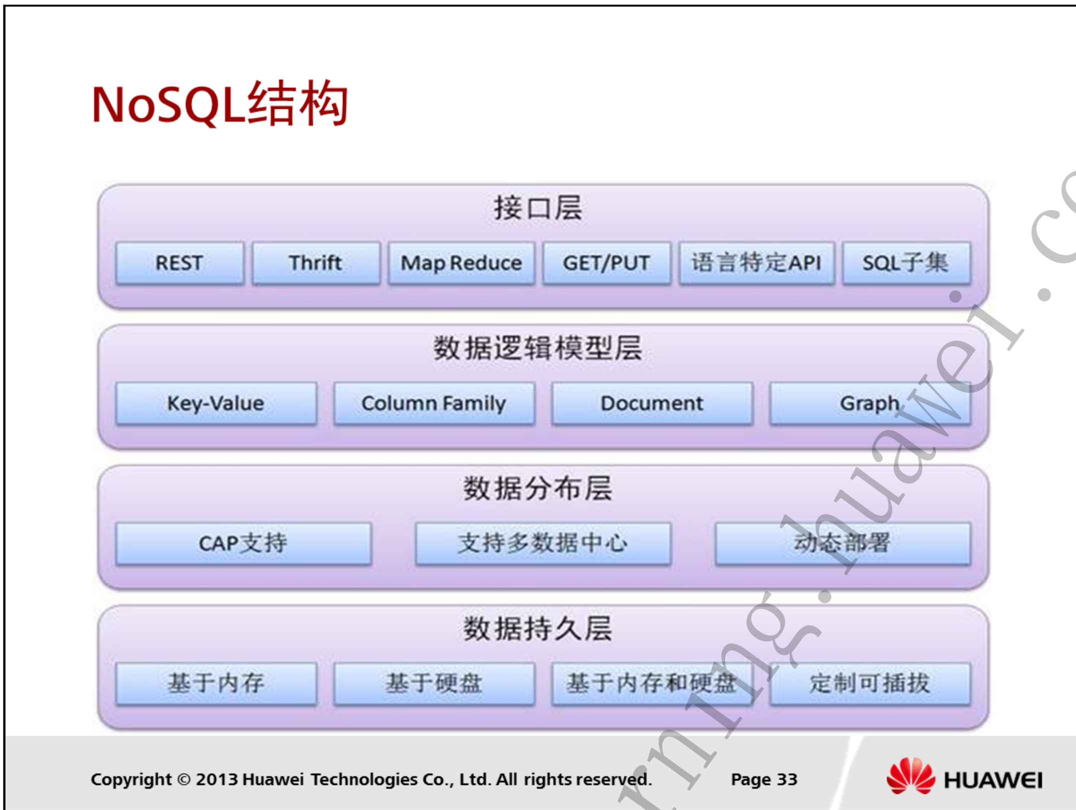


高容量、高分布式、高复杂性应用程序的需求迫使传统数据库不断扩展自己的容量极限，这些驱动传统关系型数据库采用不同的数据管理技术的6个关键因素可以概括为“SPRAIN”，即：

- 可扩展性（Scalability）——硬件价格
- 高性能（Performance）——MySQL的性能瓶颈
- 弱一致性（Relaxed consistency）——CAP理论
- 敏捷性（Agility）——持久多样性
- 复杂性（Intricacy）——海量数据
- 必然性（Necessity）——开源



- CAP原理：
  - 分布式数据系统的CAP原理的三要素：一致性(Consistency)、可用性(Availability)、分区容忍性(Partition tolerance)。
  - CAP原理：在分布式系统中，这三个要素最多只能同时实现两点，不可能三者兼顾。对于分布式数据系统，分区容忍性是基本要求。对于大多数Web应用的数据库，通常牺牲一致性而换取高可用性。



- 接口：REST (HBase，CouchDB，Riak等)，MapReduce (HBase，CouchDB，MongoDB，Hypertable等)，Get/Put (Voldemort，Scalaris等)，Thrift (HBase，Hypertable，Cassandra等)，语言特定的API(MongoDB)。
- 逻辑数据模型 分为：面向键值对的(Voldemort，Dynomite 等)，面向Column Family的(BigTable，HBase，Hypertable 等)，面向文档的(Couch DB，MongoDB等)，面向图的(Neo4j， Infogrid等)
- 数据分布模型：致性和可用性(HBase，Hypertable， MongoDB等)， 可用性和可分区性(Cassandra等)。
- 数据持久性（数据保存）：基于内存的(如Redis，Scalaris， Terrastore)，基于磁盘的(如MongoDB，Riak等)，或内存及磁盘二者的结合(如 HBase，Hypertable，Cassandra)。后一种组合方案通常最佳的选择。既能通过内存数据存储支持高性能，又能在写入足够多的数据后存储到磁盘来保证持数据存储的续性。



- 在优势方面，主要体现在下面几点：

  1、易扩展：NoSQL数据库种类繁多，但是一个共同的特点都是去掉关系数据库的关系型特性。数据之间无关系，这样就非常容易扩展。也无形之间，在架构的层面上带来了可扩展的能力。甚至有多种NoSQL之间的整合。

  2、灵活的数据模型：NoSQL无需事先为要存储的数据建立字段，随时可以存储自定义的数据格式。而在关系数据库里，增删字段是一件非常麻烦的事情。如果是非常大数据量的表，增加字段简直就是一个噩梦。

  3、高可用： NoSQL在不太影响性能的情况，就可以方便的实现高可用的架构。比如Cassandra，HBase模型，通过复制模型也能实现高可用。

  4、快速读写，大数据量，高性能：NoSQL数据库都具有非常高的读写性能，尤其在大数据量下，同样表现优秀。这得益于它的无关系性，数据库的结构简单。比如Redis，由于其逻辑简单，而且纯内存操作，使得其性能非常出色，单节点每秒可以处理超过10万次读写操作

  5、低廉的成本：这是大多数分布式数据库共有的特点，因为主要都是开源软件，没有昂贵的License成本。



- NoSQL数据库还存在着很多的不足，常见主要有下面这几个：
  1、不提供对SQL的支持：如果不支持SQL这样的工业标准，将会对用户产生一定的学习和应用迁移成本；

  2、支持的特性不够丰富：现有产品所提供的功能都比较有限，大多数NoSQL数据库都不支持事务，也不像MS SQL Server和Oracle那样能提供各种附加功能，比如BI和报表等；

  3、NoSQL未形成一定标准：各种产品层出不穷，内部混乱，现有产品的不够成熟，各种项目还需时间来检验，和关系型数据库几十年的完善不可同日而语

NoSQL分类

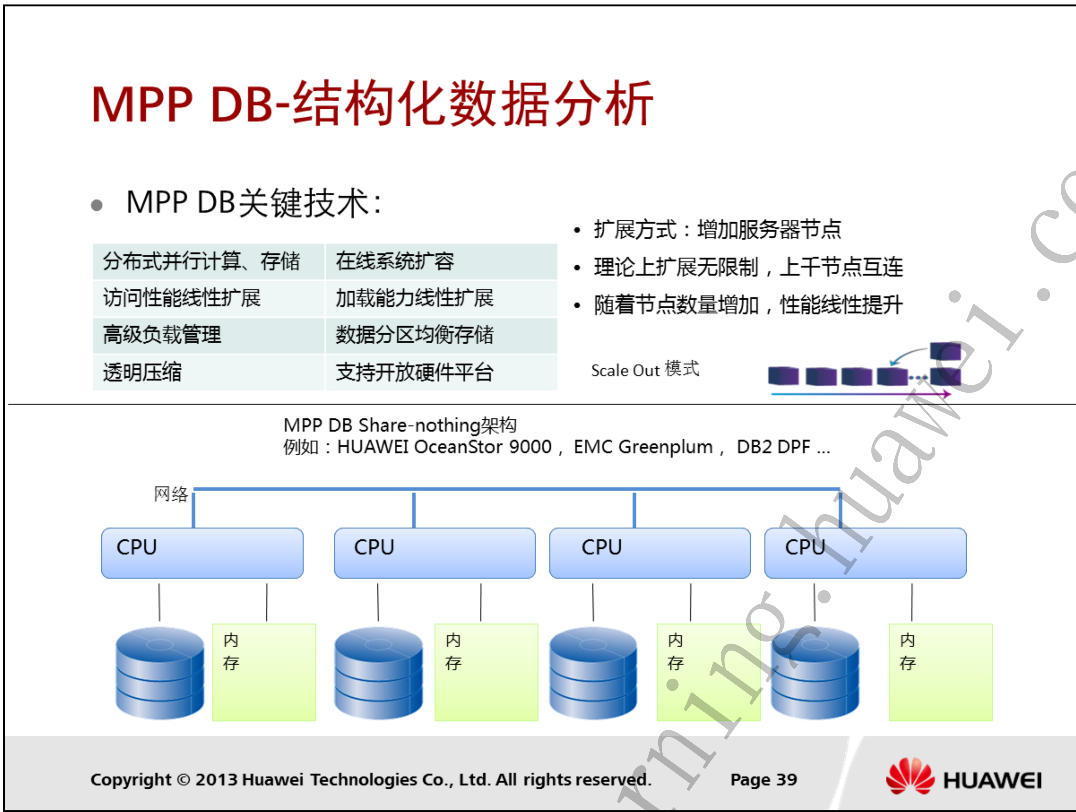
类型	部分代表	特点
列存储	Hbase/Cassandra/Hypertable	顾名思义，是按列存储数据的。最大的特点是方便存储结构化和半结构化数据，方便做数据压缩，对针对某一列或者某几列的查询有非常大的IO优势。
文档存储	MongoDB/CouchDB	文档存储一般用类似json的格式存储，存储的内容是文档型的。这样也就有有机会对某些字段建立索引，实现关系数据库的某些功能。
key-value存储	Tokyo Cabinet Tyrant/Berkeley DB/Memcache DB/Redis	可以通过key快速查询到其value。一般来说，存储不管value的格式，照单全收。（Redis包含了其他功能）
图存储	Neo4J/FlockDB/InfoGrid	图形关系的最佳存储。使用传统关系数据库来解决的话性能低下，而且设计使用不方便。
对象存储	Db4o/Versant	通过类似面向对象语言的语法操作数据库，通过对象的方式存取数据。
xml数据库	Berkeley DB XML/BaseX	高效的存储XML数据，并支持XML的内部查询语法，比如XQuery,Xpath。

NoSQL应用举例

- Google的BigTable
  - BigTable提出了一种很有趣的数据模型，它将各列数据进行排序存储。数据值按范围分布在多台机器，数据更新操作有严格的一致性保证。
- Amazon的Dynamo
  - Dynamo使用的是另外一种分布式模型。Dynamo的模型更简单，它将数据按key进行hash存储。其数据分片模型有比较强的容灾性，因此它实现的是相对松散的弱一致性：最终一致性。

NewSQL

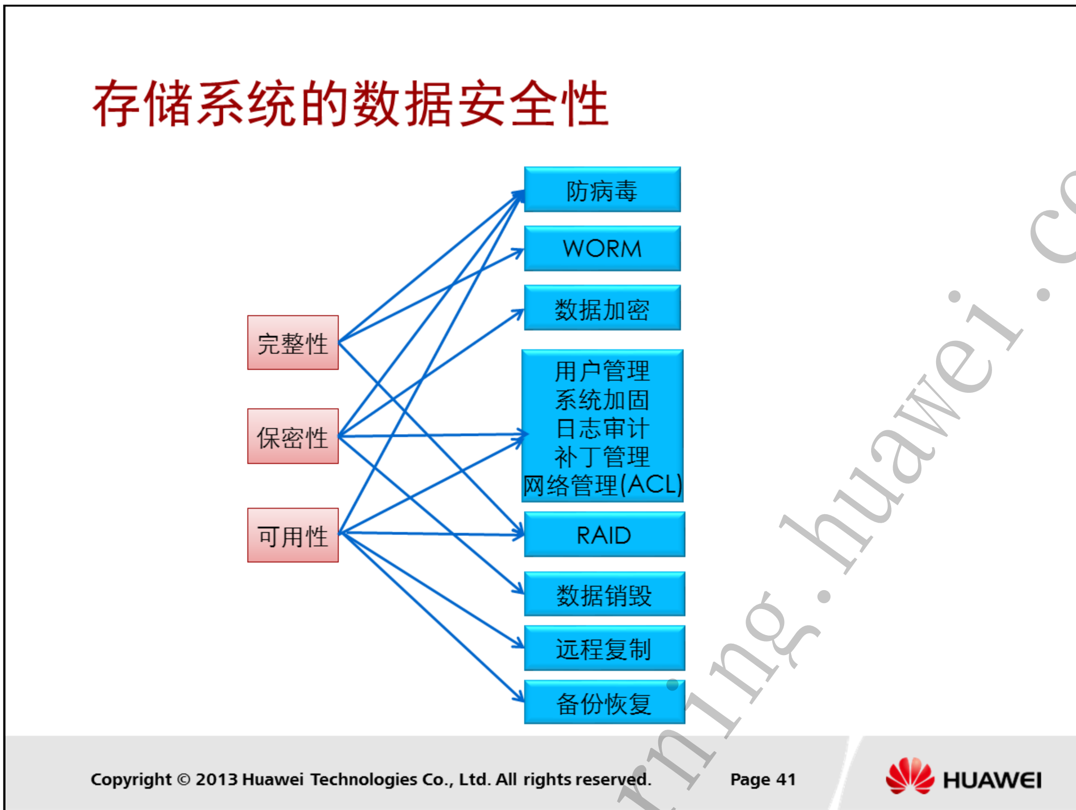
- NewSQL 是对各种新的可扩展/高性能数据库的简称。
- NewSQL数据库具有NoSQL对海量数据的存储管理能力。
- NewSQL数据库保持了传统数据库支持ACID和SQL等特性。



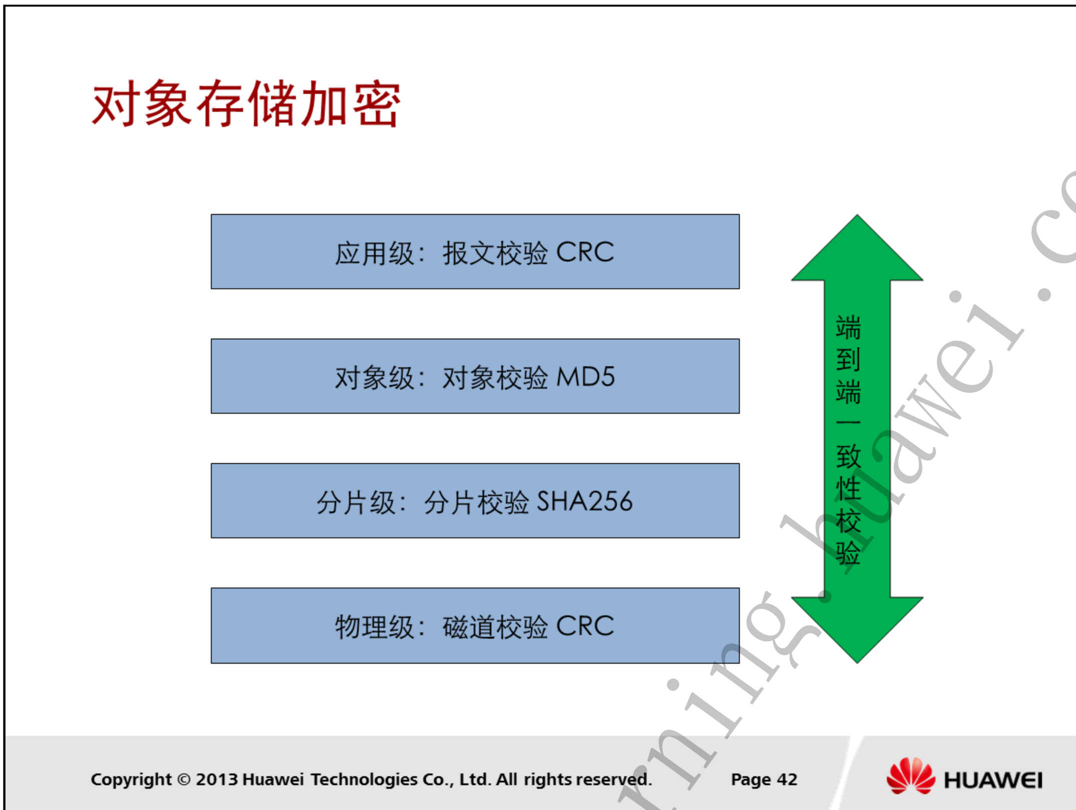
如何快速从海量数据中分析出价值数据，支撑业务决策，是大数据时代面临的一大挑战。传统的关系数据库部署难以处理TB级数据，不能很好的支持高级别的数据分析。大规模并行处理MPP数据库采取分而治之思想，让企业从容面对大数据3V（ Volume 、Value、 Velocity ）挑战。

目录

HUAWEI



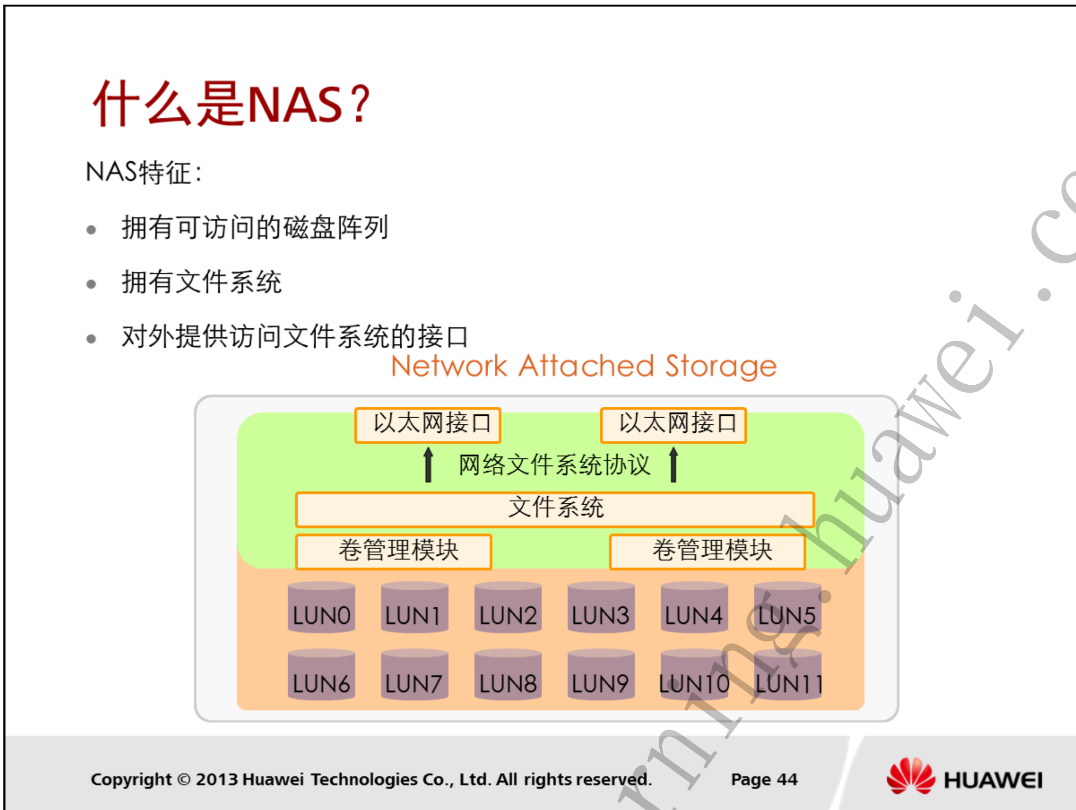
- 存储系统的数据安全性通常从数据完整性、传输和访问的保密性、数据有效读写的可用性方面考虑，有些技术或方式是存储特有，有些是通过成熟的安全技术或方案实现。
- 网络安全通常通过防火墙等安全方案或技术保障存储对外网络不受浸入。
- 应用数据的安全性通常通过防病毒软件保障 。
- 在存储数据传输过程中，数据传输协议如iSCSI、TCP/IP。
- 存储作为一种数据管理和存储的系统，也采取系统加固的一些技术和方法，如OS层面的用户管理、补丁管理等。
- 存储系统自身的功能有较高的数据安全保障技术，如RAID冗余机制提高数据存储的可靠性；远程复制和备份等特性提供数据保护综合方案。



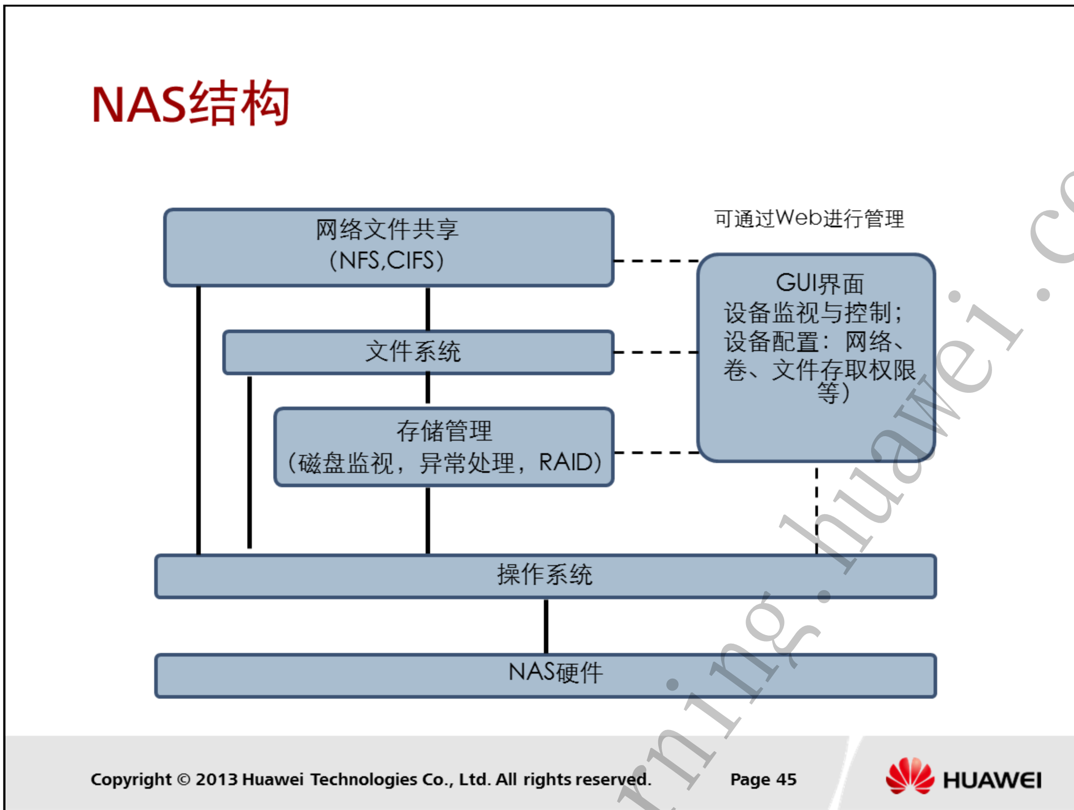
- 应用级
  - 对外通过TCP/IP和HTTP协议的校验保证系统接收和发送给用户的数据不出错。
- 对象级
  - 在接入节点处，通过MD5算法进行对象级的数据校验。
- 分片级
  - 每个分片都带有通过HMAC SHA256算法产生的校验字段，智能硬盘在接收到分片后，都会检查分片的正确性。
- 物理级
  - 每个分片连同其校验，在写入磁盘后还会增加对应磁道的校验。

目录

HUAWEI



NAS是Network Attached Storage的简称，中文称为网络附加存储。按字面简单说是连接在网络上，具备存储功能的设备，因此也称为“网络存储器”或者“网络磁盘阵列”。在NAS存储结构中，存储系统不再通过I/O总线附属于某个服务器或客户机，而直接通过网络接口与网络直接相连，由用户通过网络访问。 NAS实际上是一个带有瘦服务器（有专用OS的服务器）的存储设备，其作用类似于一个专用的文件服务器。这种专用存储服务器去掉了通用服务器原有的不适用的大多数计算功能，而仅仅提供文件系统功能。与传统以服务器为中心的存储系统相比，数据不再通过服务器内存转发，直接在客户机和存储设备间传送，服务器仅起控制管理的作用。



NAS系统软件设计的基本要求是较高的稳定性和I/O吞吐率，并能满足数据共享、数据备份、安全配置、设备管理等要求。该结构分为五个模块：操作系统、存储管理器、文件系统、网络文件共享和GUI管理模块。

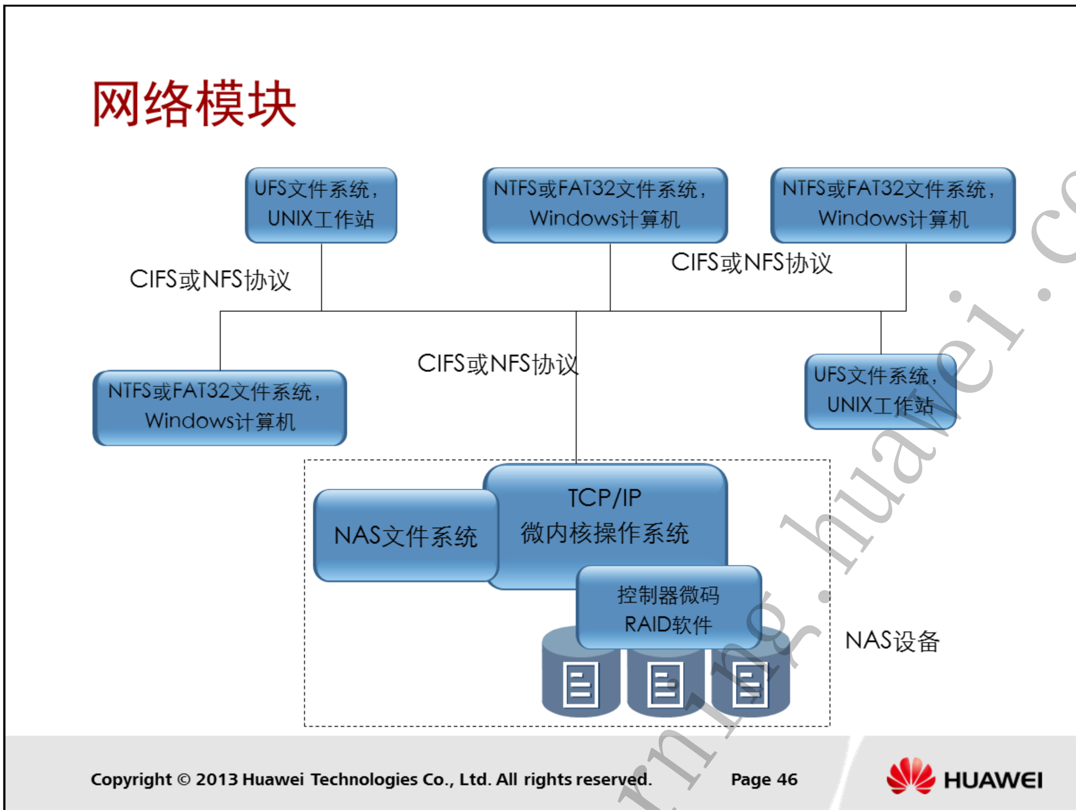
鉴于Linux、FreeBSD等免费的开放源码操作系统具有稳定、可靠、高效的优秀特性，现在大部分NAS设备都是基于此类操作系统开发的。

存储管理器的主要功能是磁盘和分区的管理，主要包括磁盘的监测与异常处理和逻辑卷的配置管理，一般应支持磁盘的热插拔、热替换等功能和RAID0、RAID1、RAID5类型的逻辑卷。存储管理器实现简化的、集中的存储管理功能，保证数据的完整性，并增强数据的可用性。

文件系统提供持久性存储和管理数据的手段，它必须是32位或以上并能支持多用户，应具备日志文件系统功能以使系统在崩溃或掉电重启后能迅速恢复文件系统的一般性和完整性，进一步提供NAS的可用性。此外，文件系统还应具有快照功能。快照不仅能恢复被用户错误修改或删除的文件，而且能实现备份窗口为零的文件系统活备份。

网络文件共享一般支持以下一些文件传输和共享协议，如FTP和HTTP协议、Unix系统的NFS、Windows系统的CIFS、Novell系统的NCP（Novell core protocol）、Apple系统的AFP（appletalk file protocol）等，因此NAS设备具有较好的协议独立性。

GUI管理提供给系统管理员一个友好的界面，使之仅通过web浏览器操作就能远程监视和管理NAS设备的系统参数，如：网络配置、用户与组管理、卷以及文件共享权限等。



NAS设备中所包含的标准文件系统可以对公用互联网文件系统（CIFS）或是网络文件系统（NFS）提供支持，也有可能同时支持二者。在许多情况下，它都使用标准的网络文件系统来作为NAS专用文件系统的接口。大多数NAS设备需要用这种方式来管理其自身的存储资源。

NFS（网络文件系统）是Unix系统间实现磁盘文件共享的一种方法，支持应用程序在客户端通过网络存取位于服务器磁盘中数据的一种文件系统协议。其实它包括许多种协议，最简单的网络文件系统是网络逻辑磁盘，即客户端的文件系统通过网络操作位于远端的逻辑磁盘。现一般在Unix主机之间采用Sun开发的NFS（Sun），它能够在所有Unix系统之间实现文件数据的互访，逐渐成为主机间共享资源的一个标准。

CIFS是由微软开发的，用于连接Windows客户机和服务器。经过Unix服务器厂商的重新开发后，它可以用于连接Windows客户机和Unix服务器，执行文件共享和打印等任务。它最早的由来是NetBIOS，这是微软开发的在局域网内实现基于Windows名称资源共享的API。之后，产生了基于NetBIOS的NetBEUI协议和NBT(NetBIOS OVER TCP/IP)协议。NBT协议进一步发展为SMB（Server Message Block Potocol）和CIFS（Common Internet File System，通用互联网文件系统）协议。其中，CIFS用于Windows系统，而SMB广泛用于Unix和Linux，两者可以互通。SMB协议还被称作LanManager协议。CIFS支持与SMB的服务器通信而实现共享。微软操作系统家族和几乎所有Unix服务器都支持SMB协议/SAMBA软件包。

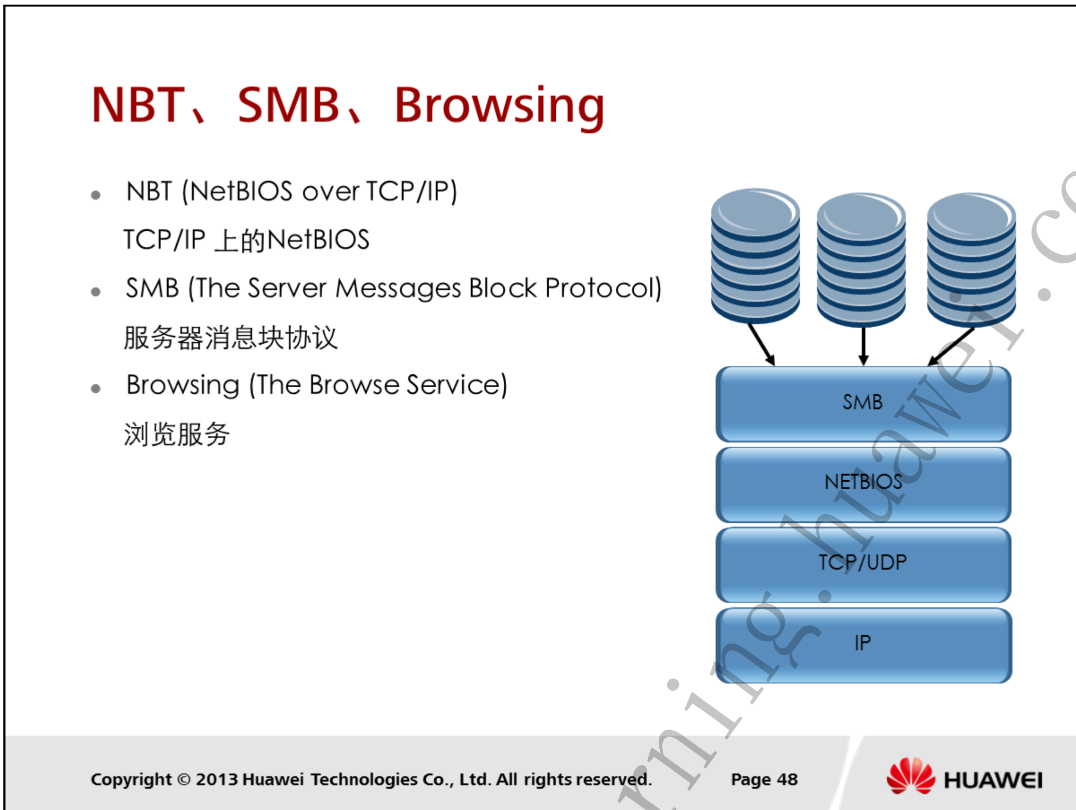
CIFS和NFS的对比。CIFS面向网络连接的共享协议，对网络传输的可靠性要求高，常使用TCP；NFS是独 立于传输的，可使用TCP或UDP。

CIFS是什么

- 定义
  - CIFS (Common Internet File System)，通用Internet文件系统，一个新提出的协议，它使程序可以访问远程Internet计算机上的文件并要求此计算机的服务。
- CIFS是公共的或开放的SMB协议版本，包含主要模块有：
  - NBT
  - SMB
  - Browsing

Page 47
HUAWEI

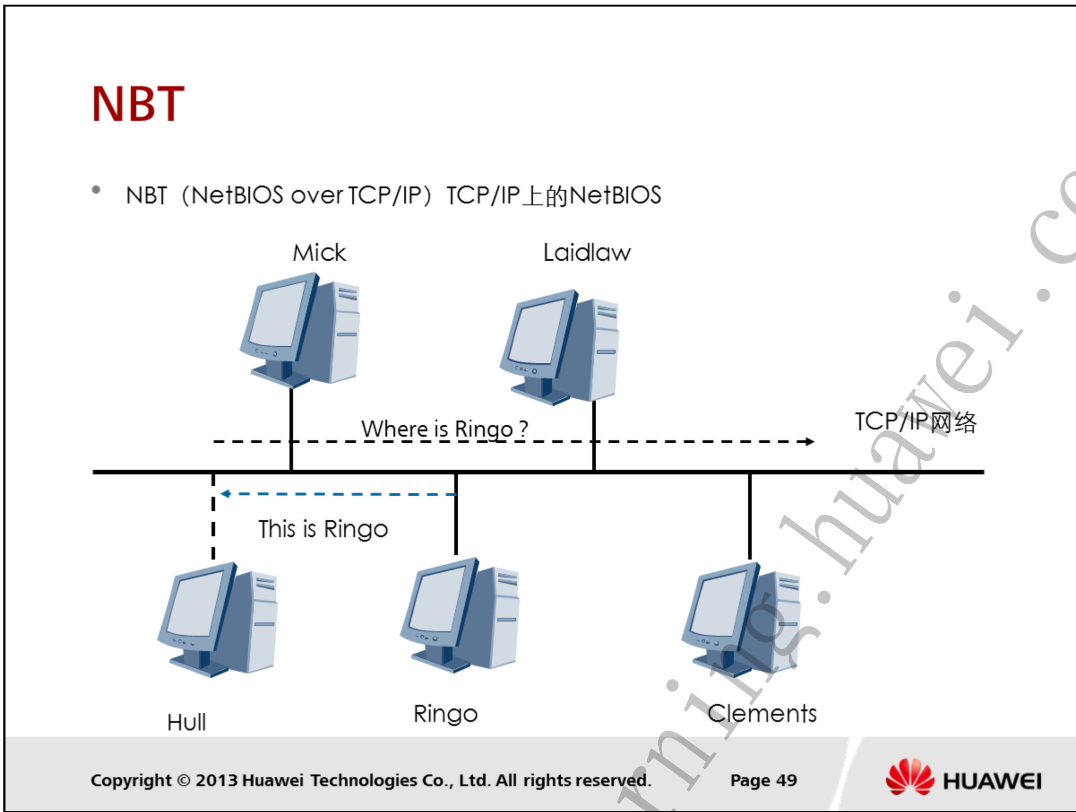
CIFS使用客户/服务器模式，客户程序请求远在服务器上的服务器程序为它提供服务，服务器获得请求并返回响应。



TCP/IP上面的NetBIOS也叫做NetBT或者NBT。

在OSI的7层参考模型里面，NetBIOS是一个会话层（第5层）编程接口。在DOS和它的后代，应用层通过填写一个叫做网络控制块(NCB)然后发信号的方式和NetBIOS沟通。NCB是用来在应用之间通过底层协议栈传送命令和消息。

NetBIOS编程接口是DOS和它的家族所特有的。Unix和其它操作系统不需要实行NetBIOS编程接口，也就没有程序需要使用它。

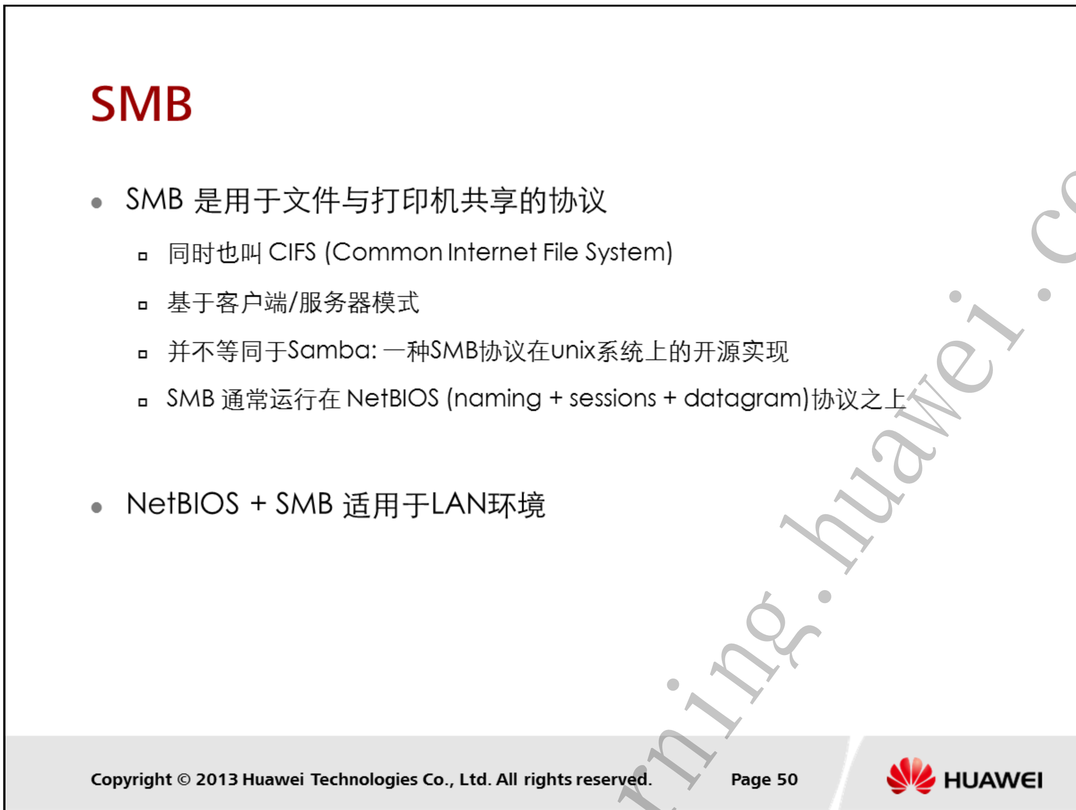


NetBIOS局域网架构是很简单的。没有路由器，没有交换机——只是一堆用（虚拟的）线连接在一起的节点。不必区分硬件地址、网络地址，甚至象在IP里面那样需要区分端口号。取而代之的是，通信节点用一个16字节的字符串，即“NetBIOS名字”来标识。

每个节点保留一个它已经注册的名字的列表（就是它拥有的名字）。当发送一个消息的时候，第一步就是发送一个IP广播查询，叫做“名字查询请求”。如果在IP局域网有一个机器拥有被查询的名字，它发送一个“名字查询应答”的答复来应答。

要实现NetBIOS局域网的仿真必须支持的四个基本服务：

- the Name Service 名字服务
- the Datagram Service 数据报服务
- the Session Service 会话服务
- the Name Service 名字服务



服务器消息块协议（SMB），一种客户机/服务器、请求/响应协议。通过 SMB 协议，客户端应用程序可以在各种网络环境下读、写服务器上的文件，以及对服务器程序提出服务请求。

服务器信息块（SMB）协议，用于在计算机间共享文件、打印机、串口等。SMB协议可以用在因特网的TCP/IP协议之上，也可以用在其它网络协议如IPX和NetBEUI 之上。

在 TCP/IP 环境下，客户机通过 NetBIOS over TCP/IP（或 SPX/IPX或 NetBEUI/TCP ）连接服务器。一旦连接成功，客户机可发送 SMB 命令到服务器上，从而客户机能够访问共享目录、打开文件、读写文件，以及一切在文件系统上能做的所有事情。

目前主流的Windows 操作系统已包括了客户机和服务器 SMB 协议支持。对于 UNIX 系统，可使用一种称为 Samba 的共享软件。

SMB 定义了两级安全保护：

- 共享级保护（Share Level Protection）应用于服务器共享目录级。每个共享目录都需要提供一个访问口令。只有口令通过，客户机才能访问所有共享文件。
- 用户级保护（Usr Level Protection）应用于共享目录中的单独文件，基于用户访问权限。每个用户（客户机）必须登录服务器并且获得服务器的认证许可。一旦认证通过，客户即会获得一个 UID ,在以后客户机访问服务器的过程都需要使用该 UID 。

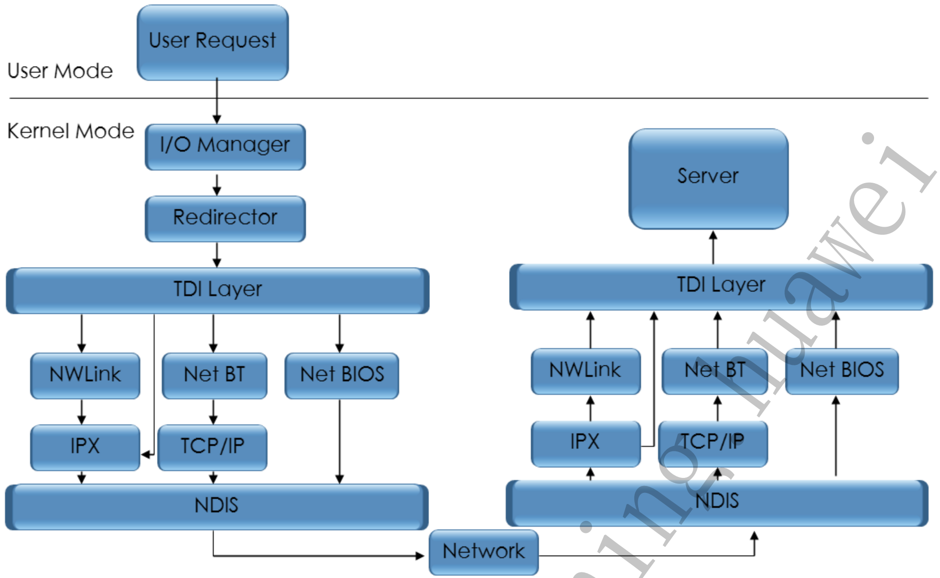


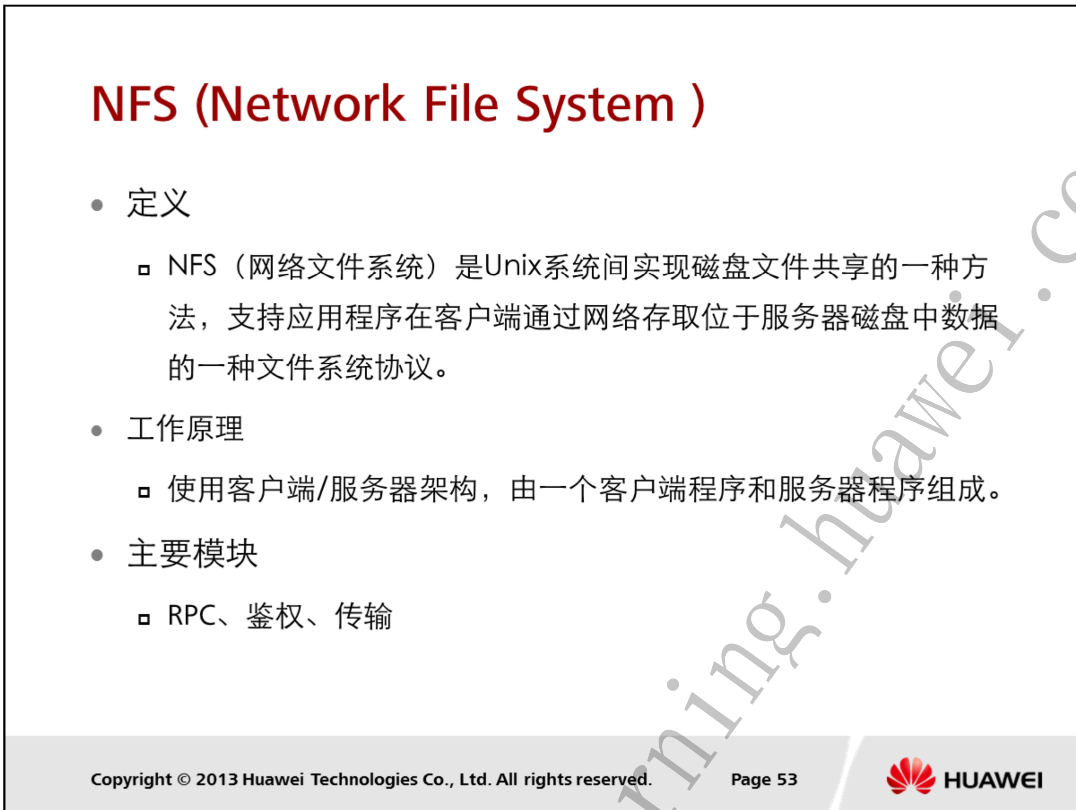
Browsing（The Browse Service ）浏览服务，微软定义的一种完成网络资源查找和定位机制，目的是为了帮助用户找到网络资源。

浏览列表（Browsing List）：当通过网上邻居窗口打开整个网络时，将看到一个工作组列表，再打开某个工作组，将看到里面的计算机列表，就是我们所说的 Browsing List。在微软网络中，用户可以在浏览列表里看到整个网络上所有的计算机。

浏览列表其实是从工作组中的某一台计算机（一般叫主浏览器 ）中获得 。在一个工作组中只有一台计算机充当主浏览器的角色，有一到两台计算机充当备份浏览器的角色 。如果不指定，主浏览器、备份浏览器身份都是由网络中的计算机随机得到。如果网络中的机器都是Win98系统，则由第一台打开并启用文件及打印共享的计算机充当主浏览器 。如果主浏览器关机，就要选举主浏览器了 。

CIFS 消息流

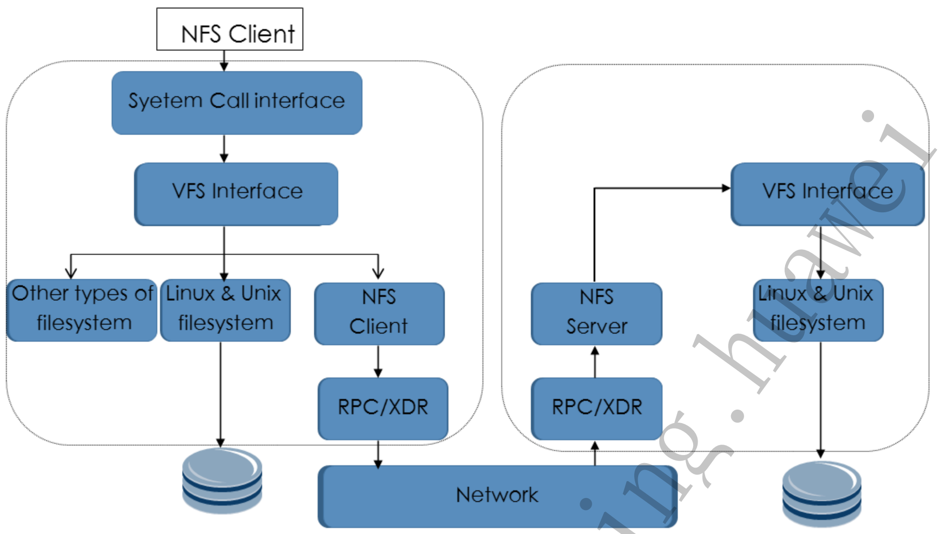

User Request
User Mode
Kernel Mode
I/O Manager
Redirector
TDI Layer
NWLink
Net BT
Net BIOS
IPX
TCP/IP
NDIS
Server
TDI Layer
NWLink
Net BT
Net BIOS
IPX
TCP/IP
NDIS
Network



NFS (Network File System，网络文件系统)是当前主流异构平台共享文件系统之一。主要应用在UNIX环境下。 最早是由SUN microsystem开发，现在能够支持在不同类型的系统之间通过网络进行文件共享，广泛应用在FreeBSD、SCO、Solaris等等异构操作系统平台， 允许一个系统在网络上与它人共享目录和文件。通过使用NFS，用户和程序可以象访问本地文件一样访问远端系统上的文件，使得每个计算机的节点能够像使用本地资源一样方便地使用网上资源。换言之，NFS 可用于不同类型计算机、操作系统、网络架构和传输协议运行环境中的网络文件远程访问和共享。

NFS的工作原理是使用客户端/服务器架构，由一个客户端程序和服务器程序组成。服务器程序向其它计算机提供对文件系统的访问，其过程就叫做“输出”。NFS 客户端程序对共享文件系统进行访问时，把它们从 NFS 服务器中“输送”出来。文件通常以“块”为单位进行传输. 其尺寸是 8K (虽然它可能会将操作分成更小尺寸的分片).NFS 传输协议用于服务器和客户机之间文件访问和共享的通信，从而使客户机远程地访问保存在存储设备上的数据。

NFS消息流

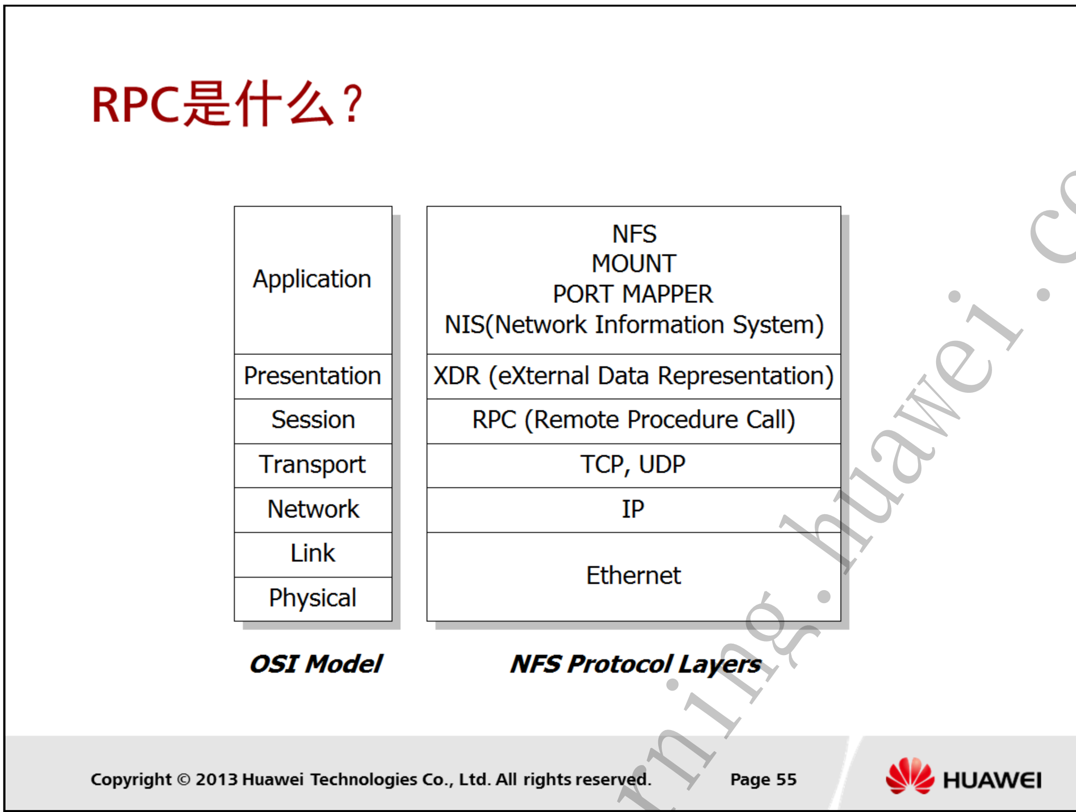


 Page 54 HUAWEI

NFS工作流程，当NFS客户端发送请求访问远端NFS服务器上的文件时，内核会通过VFS层，将请求通过RPC发送到远端的NFS服务器，RPC可以通过TCP/UDP来进行传输，然后NFS服务器在2049端口接收到NFS客户端的RPC请求，并将其转发到NFS服务器的本地文件系统。

NFS服务器在请求完成后，会将结果通过RPC发送回执到NFS客户端，其中有两点需要注意:

- NFS服务器发送回执到NFS客户端需要一定的时间，此时NFS服务器同时仍然能响应其他的NFS客户端请求，此时服务器上存在多个NFSD的进程。
- 在NFS客户端也存在相同的情况，NFS客户端会存在多个biod进程，用于发送NFS请求。



- RPC（Remote Procedure Call）
  - NFS本身不提供信息传输的协议和功能，而是使用了一些其它的传输协议。而这些传输协议用到这个RPC功能的。可以说NFS本身就是使用RPC的一个程序。或者说NFS也是一个RPC SERVER.所以只要用到NFS的地方都要启动RPC服务。这样SERVER和CLIENT才能通过RPC来实现PROGRAM PORT的对应。可以这么理解RPC和NFS的关系：NFS是一个文件系统，而RPC是负责负责信息的传输。
  - 网络七层协议与NFS协议的对应关系：NFS是处于应用层的协议，NFS的相关模块为mount，portmapper模块，mount进程可以对客户端的请求进行认证，portmaaper提供了客户端要连接的端口，NIS提供了一种域的认证方式，NFS及其相关协议都需要使用到RPC，所有的NFS请求都需要通过RPC来完成，XDR是RPC的一种数据编码格式。RPC可以使用TCP/UDP进行传输。

传输方式

- 使用UDP协议
  - UDP传输在NFS局域网的应用中传输速度快
  - UDP协议传输的开销小
- 使用TCP协议
  - 可靠性高，有效的阻塞控制
  - 客户端和服务器端都保留TCP连接的状态
  - 服务器崩溃时，客户端只需要打开一个新的TCP连接
  - 客户端崩溃时，服务器端在新的TCP连接到来时，关闭原来的TCP连接

 Page 56 HUAWEI

不论NFS使用UDP还是TCP，NFS连接都是无状态的。

应用环境对比、网络特性对比

特性	NFS	CIFS
传输协议	TCP/UDP	TCP
连接状态	无状态	有状态
故障恢复	自动恢复	无
传输包大小	8K/32K	64K
文件格式	UNIX类型	WIN类型

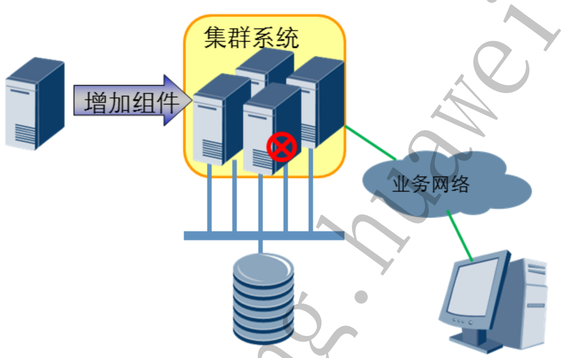
HUAWEI

目录

HUAWEI

集群技术

- 定义：
  - 集群是由一组相互独立的服务器组成，对外表现为单一服务器，提供高可靠性服务。
- 特点：
  - 统一命名
  - 高可靠性
  - 性能扩展
  - 共享数据空间



 Page 59 HUAWEI

- 集群技术定义：
  - 一组相互独立的服务器在网络中表现为单一的系统，并以单一系统的模式加以管理。此单一系统为客户工作站提供高可靠性的服务。
- 集群技术特点：
  - 统一命名：大多数模式下，集群中所有的计算机拥有一个共同的名称，集群内任一系统上运行的服务可被所有的网络客户所使用。
  - 高可靠性：集群必须可以协调管理各分离的组件的错误和失败，集群内单一系统的失败由其他集群成员来弥补，对客户是不可见的。集群内部各节点服务器通过一内部局域网相互通讯。当一台节点服务器发生故障时，这台服务器上所运行的应用程序将在另一节点服务器上被自动接管。当一个应用服务发生故障时，应用服务将被重新启动或被另一台服务器接管。当以上的任一故障发生时，客户都将能很快连接到新的应用服务上。
  - 性能扩展：并可透明地向集群中加入组件，提升系统性能。
  - 共享数据空间：一个集群包含多台(至少二台)拥有共享数据存储空间的服务器。任何一台服务器运行一个应用时，应用数据被存储在共享的数据空间内。每台服务器的操作系统和应用程序文件存储在其各自的本地储存空间上。

集群NAS存储系统概念与特点

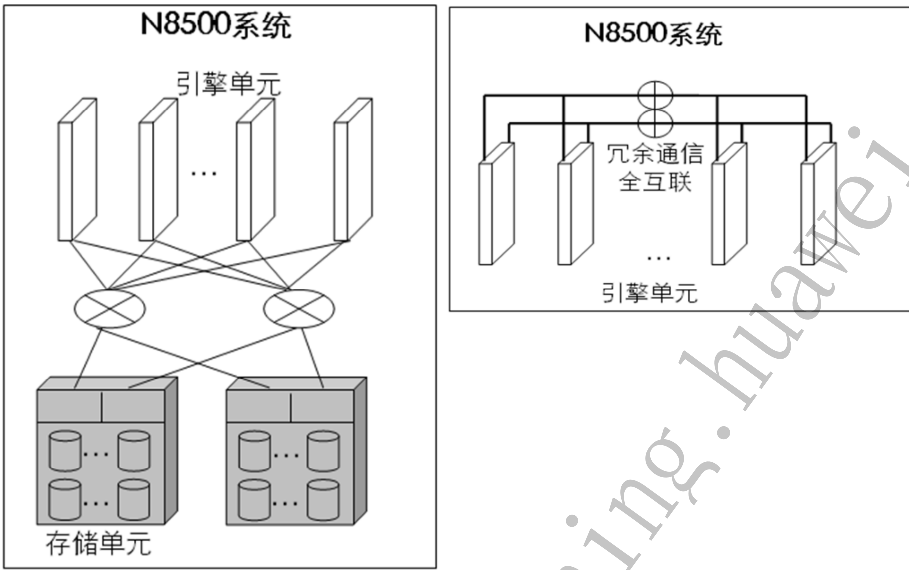
- 集群NAS（Network Attached Storage）系统
  - 概念：集群NAS存储系统是指将集群技术应用到NAS存储系统中，是为应对非结构化数据应用暴涨，提升NAS文件共享服务的性能和高可靠性。
  - 集群NAS存储系统与传统NAS存储系统的对比：

	传统NAS存储系统	集群NAS存储系统
可靠性	单控模式：单点故障；双控模式：故障切换导致业务中断，链路切换时间长。	故障切换快速；双控模式采用Active-Active模式，集群单个节点故障不影响业务。
高性能	多套设备提高性能将会带来数据冗余，数据不一致等问题。	引擎架构可扩展；共享存储设备；文件系统位于存储侧，读取速度快。
可扩展性	固定引擎架构；频繁更新及添加导致的系统停机开始影响业务	在多层面提供扩展优势：NAS机头，存储容量；支持在线扩展

 Page 60 HUAWEI

- 集群NAS存储系统的优势：
  - 统一存储：NAS服务器和存储分离。同时支持NAS和SAN应用。可在通过NAS服务器提供文件服务的同时，提供直接访问后端存储。
  - 高性能：文件系统位于存储侧而不再是服务器侧，从而回避了文件读写过程中数据在链路中反复传递的过程，加速了文件的读取速度。
  - NAS专用的操作系统针对数据的读取操作采取了更专业的优化设置，减少了磁盘寻道时间；专用的文件系统更有利于提高文件的读取效率；提供多路径冗余，可实现应用链负载均衡；提供分级存储功能，可实现数据自动迁移、快速故障切换、迁移和切换对应用几乎没有影响。

集群NAS数据通道互联

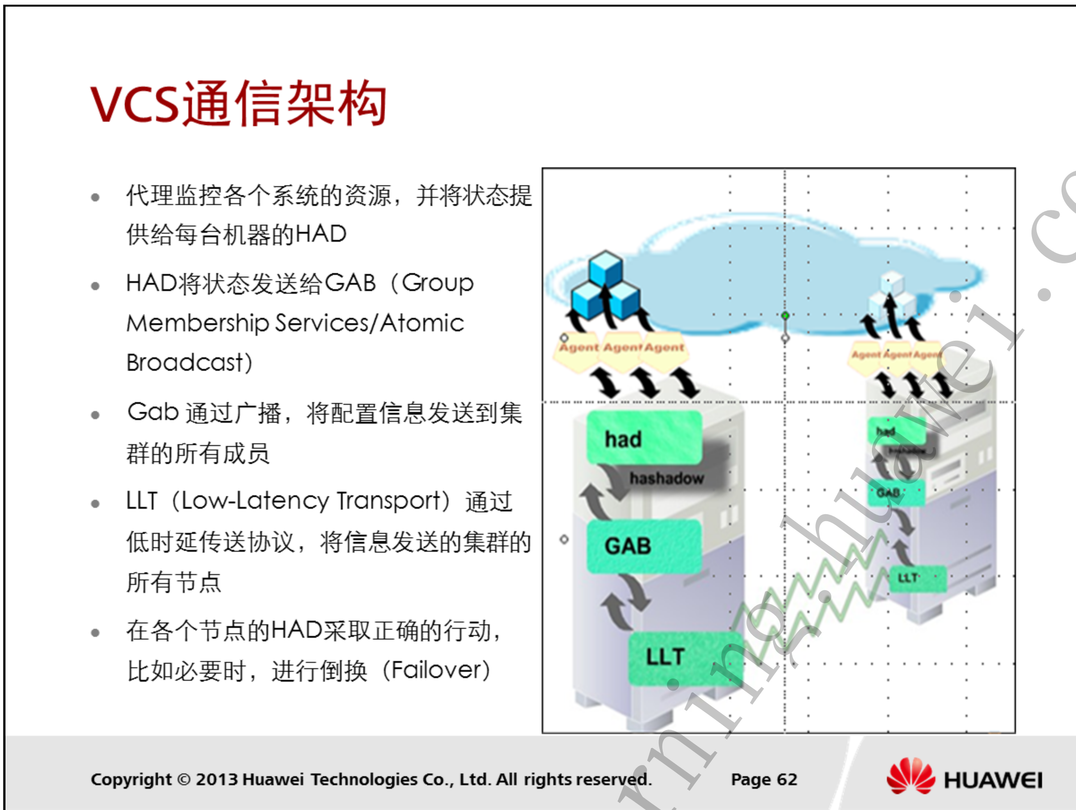


 Page 61 HUAWEI

集群NAS系统分为引擎单元和存储单元，存储单元处理数据块的存储，引擎单元提供集群文件系统的逻辑功能。引擎单元和存储单元通过直连全互联，或者交换网络实现全互联，从任何一个引擎节点都能够访问到任何一个存储单元，从而保证数据通道的畅通无阻，并且提供冗余路径实现高可靠性。

对于N8500系统的典型配置，两个引擎节点连接到存储单元，每个节点都有路径连接到阵列的双控端口中，从而实现全互联，并且提高冗余链路；对于N8500系统的典型配置，引擎节点通过冗余的交换机连接到存储单元，每个存储单元也分别连接到冗余交换机，从而形成全冗余的组网。

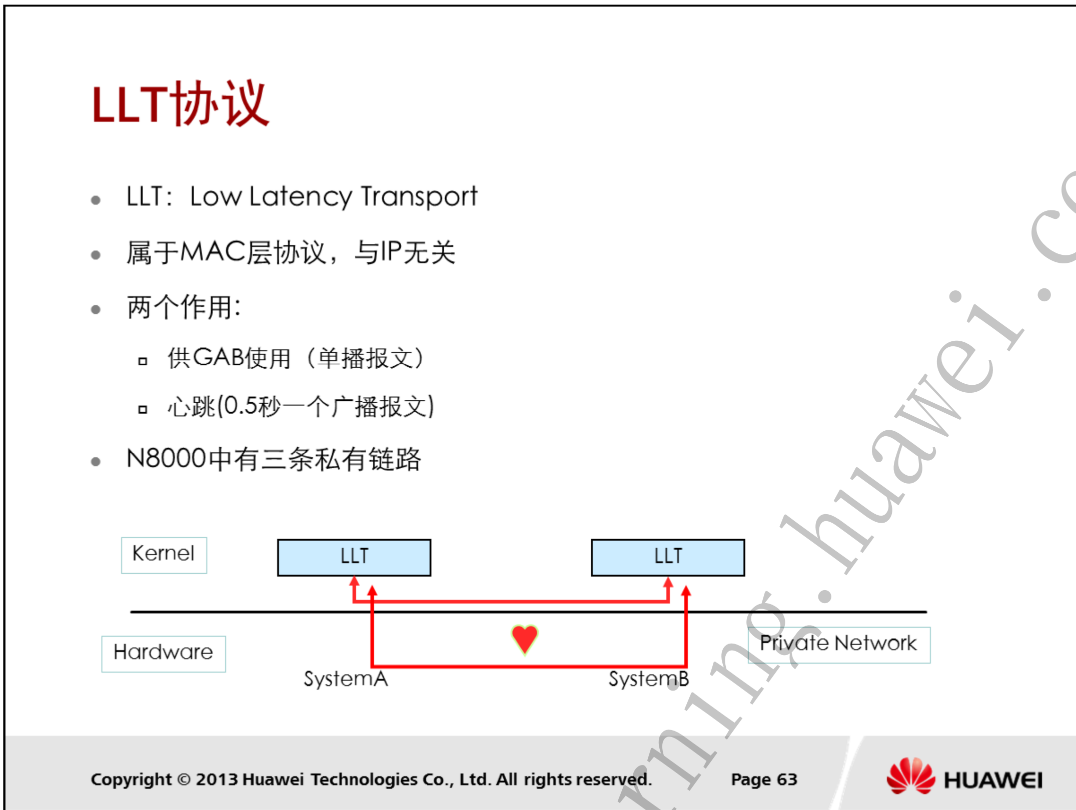
N8500系统的2引擎节点典型配置下，每个节点提供2条链路到另外一个节点，从而即使一条链路故障时，也能保证通信链路的可靠性。N8500系统的多引擎节点，则是连接到2个交换机上面，从而提供冗余通信平面，保证可靠性。



N8500系统的GLM集群技术是构建在LLT(Low Latency Transport)/GAB(Group Membership and Atomic Broadcast)底层通信基础上，该通信层直接构建在以太网二层协议上，而不是TCP/IP层上。考虑到集群通信主要传送的锁信息，通常帧都比较小，而TCP/IP包会带来额外的大帧头开销；同时集群通信主要在私有网络内传递，对TCP/IP在广域网的超时重传、滑动窗口没有太多需求。

所以，LLT针对集群锁通信需求，直接利用以太网帧传送包，从而达到提高利用率的目的，并且设计自己的流量控制、超时重传、以及重复帧处理，实现高效传输。

同时，针对集群多引擎节点通信的原子要求，设计了GAB模块，它保证通信的原子性；也就是说，保证在集群多引擎节点中，消息要么被所有引擎节点受到并确认，要么所有引擎节点都丢弃该消息，从而不会出现部分确认消息、部分丢弃消息的中间状态。



- N8000V2R1_01:~ # cat /etc/llttab

  set-node N8000V2R1_01

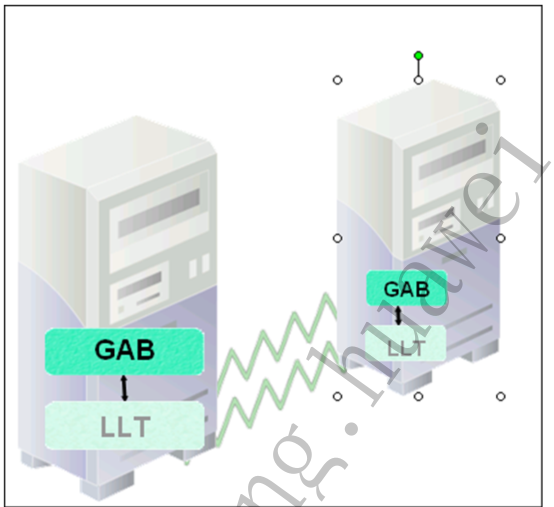
  set-cluster 130

  link priveth0 eth-00:22:a1:0a:7b:63 - ether - -

  link priveth1 eth-00:22:a1:0a:7b:61 - ether - -

GAB

- 原子广播，实际上是点到点的单播
- 使用LLT协议发送数据
- 管理集群成员：gabconfig -a



 Page 64 HUAWEI

Group Membership Services/Atomic Broadcast (GAB) 协议负责集群成员集通信和集群通信。

- 集群成员集：GAB经由LLT从每个节点接收有关心跳状态的输入来维护集群成员集。当系统不再从对等端接收心跳时（16S），会将对等端标记为DOWN，并将其从集群中排除。在VCS中，成员集是参与集群的系统的集合。
- 集群通信：GAB的第二项功能是可靠的集群通信。GAB可确保将点到点消息和广播消息发送到所有节点。VCS引擎使用专用IOCTL（由GAB提供）来判定GAB是否处于活动状态

在N8000中的六个gab成员，正常状态和部分项的启动命令如下：

Port a gen bab407 membership 0123 heartbeat membership /etc/init.d/llt start

Port b gen bab409 membership 0123 I/O fencing membership /etc/init.d/vxfen start

Port f gen bab412 membership 0123 CFS

Port h gen bab40d membership 0123 VCS hastart

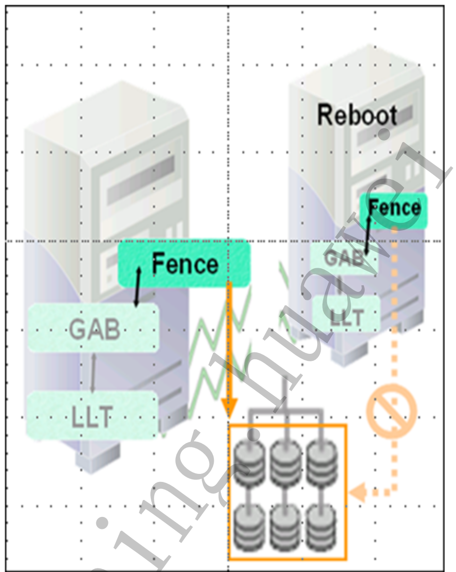
Port v gen bab40d membership 0123 CVM

Port w gen bab40f membership 0123 Cluster Volume Manager daemons

fencing

Fencing 功能

- 监控GAB，并检测集群成员的变化
- 保证集群成员的单一视图，防止集群脑裂
- 防止从多个节点，在没有单一
- 集群控制下，用卷管理器访问共享存储



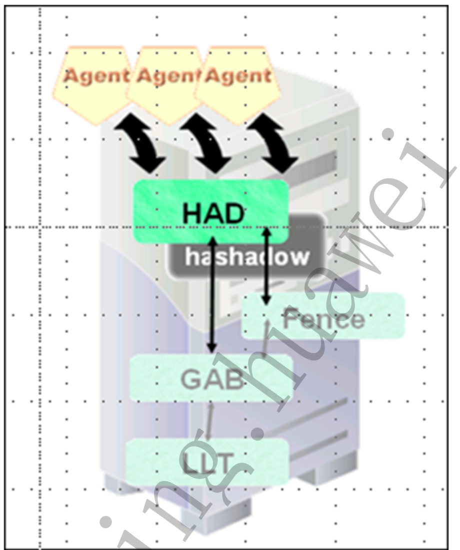
HUAWEI

1.开启fencing需要阵列支持scsi3

2.在N8000中，需要3个盘作为仲裁盘，建议磁盘来自同一套阵列的不同raid组，每个协调盘大小为100，不能使用主机lun id为0。

HAD

- VCS的引擎，就是HAD（High availability daemon），其功能为：
  - 在集群的每个主机系统上都有该进程运行
  - 对所有的集群资源，维护配置和状态信息
  - 管理所有的代理
  - 用hashadow监控进程HAD，在HAD进程故障时，能够把它重新启动起来



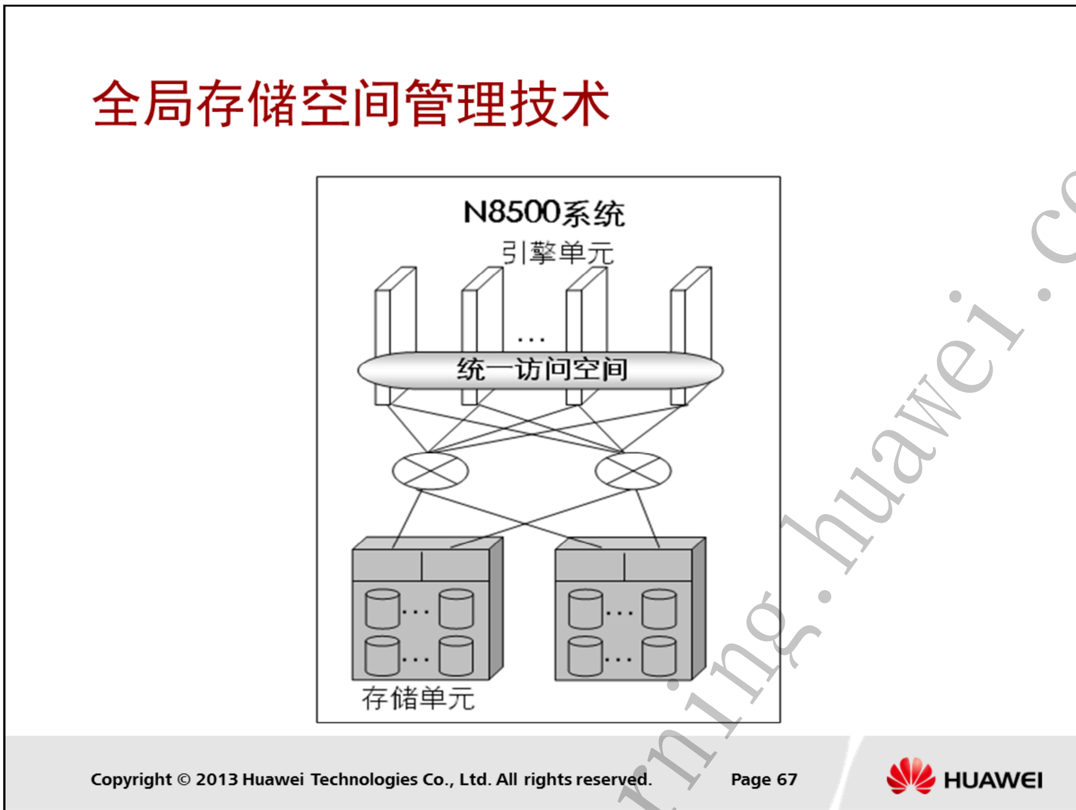
 Page 66 HUAWEI

1.使用ps –ef | grep had,可以看到两个进程

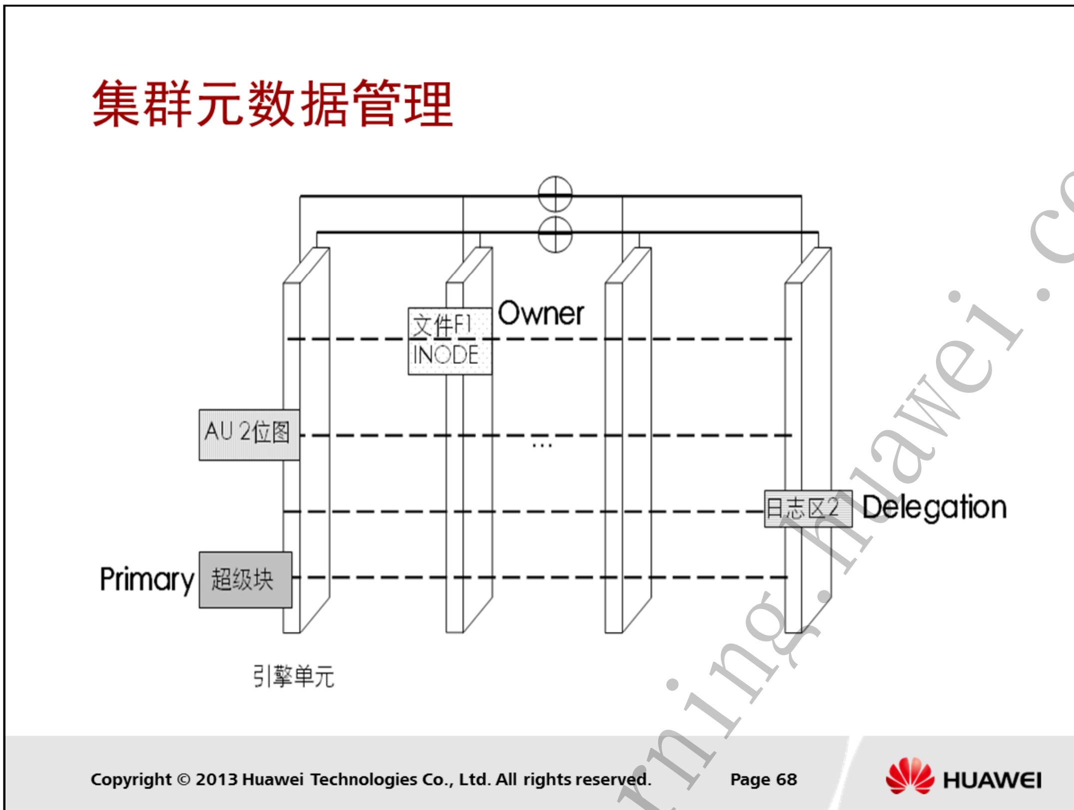
root 13317 1 0 Oct22 ? 00:01:16 /opt/VRTSvcs/bin/had

root 13347 1 0 Oct22 ? 00:00:00 /opt/VRTSvcs/bin/hashadow

2.使用hastart命令可以自动开启had和hashadow



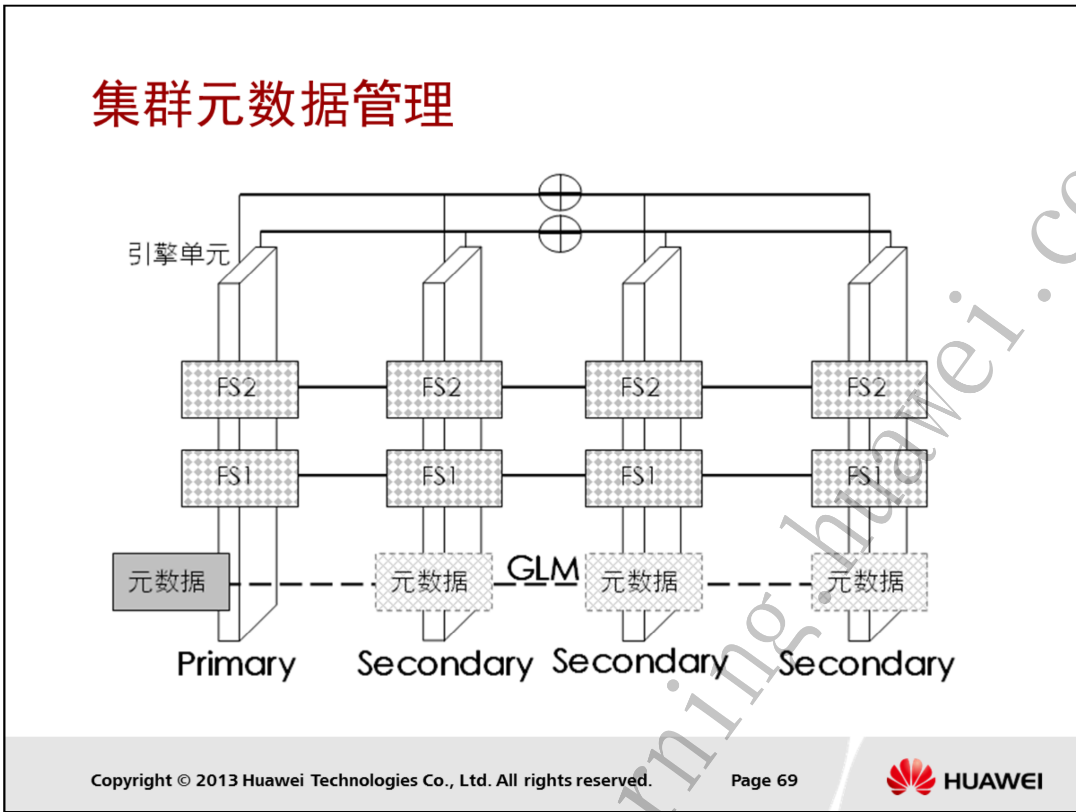
集群NAS系统在数据通道和通信通道实现全互联的基础上，采用集群卷管理技术，提供了全局的存储空间管理。也就是说，在集群中的任何节点能够以相同的卷设备名字访问同一存储空间，保证块设备访问的一致性。



N8500系统中，元数据分为了不同的类型，比如：超级块、日志区、分配单元(AU, Allocation Unit)、INODE。这些元数据具有不同的属性，分别对不同的数据管理负责，比如，超级块对整个集群文件系统的数据负责，而INODE则是针对单独的那个文件负责。

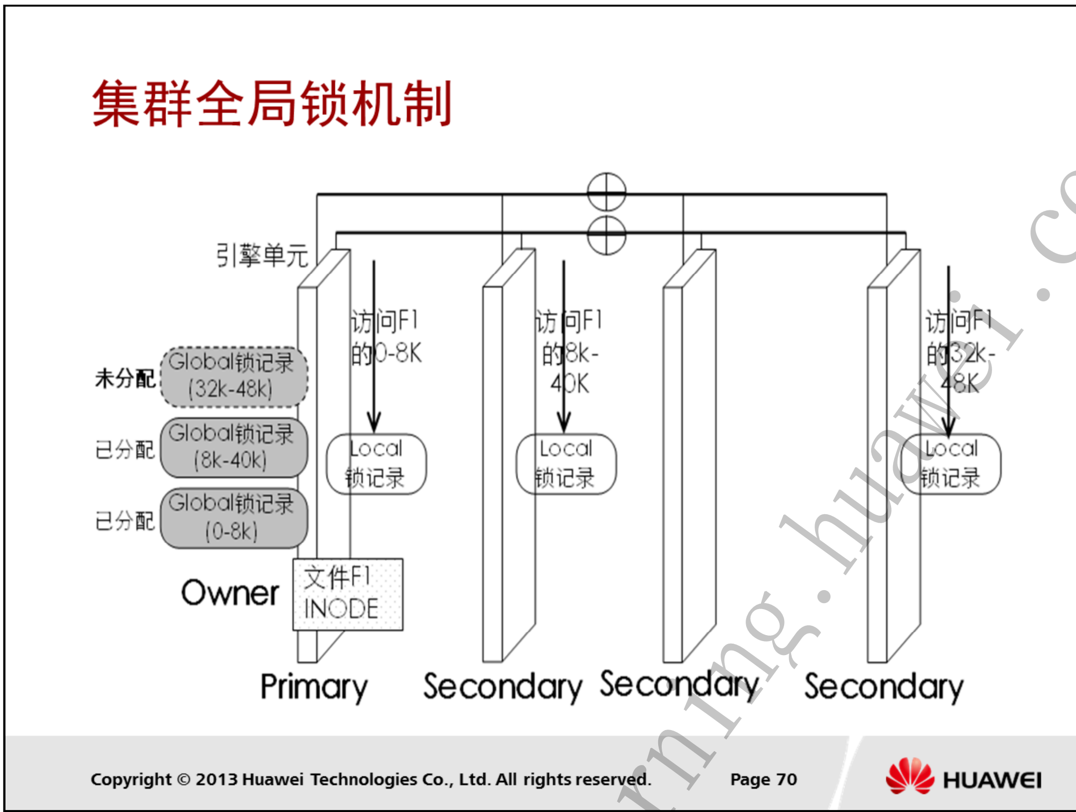
在元数据分类的基础上，N8500系统的每个引擎节点就可以担当不同元数据处理的责任主体。如图所示，第一个引擎节点负责超级块的更新修改，其他节点需要对元数据的处理，都把请求发送给他；第二个引擎节点负责文件F1的INODE元数据更新修改，其它节点需要操作F1文件，都需要把请求发送给该引擎节点。

同时，在任何引擎节点故障后，都可以把该节点负责的元数据处理转移到其他任何引擎节点上，从而实现对称式的元数据处理，实现元数据处理的负载均衡，提高元数据处理的效率。



N8500系统采用对称式的设计架构，每个节点都可以参与处理元数据，从而很容易形成全局名字空间，让系统的每个处理节点都能够读写相同的文件系统，实现节点之间的并发访问，如图所示。

根据元数据的分类、分片，让每个节点都能参与元数据的处理，减少节点之间转发元数据/数据请求的通信量，实现负载均衡。

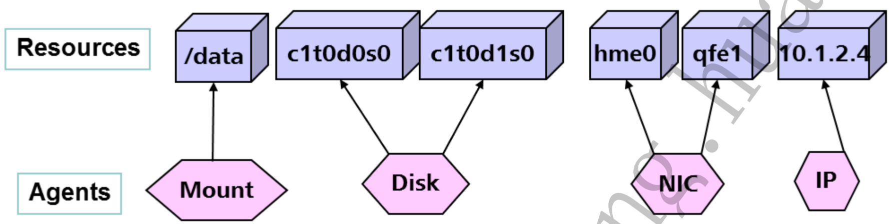


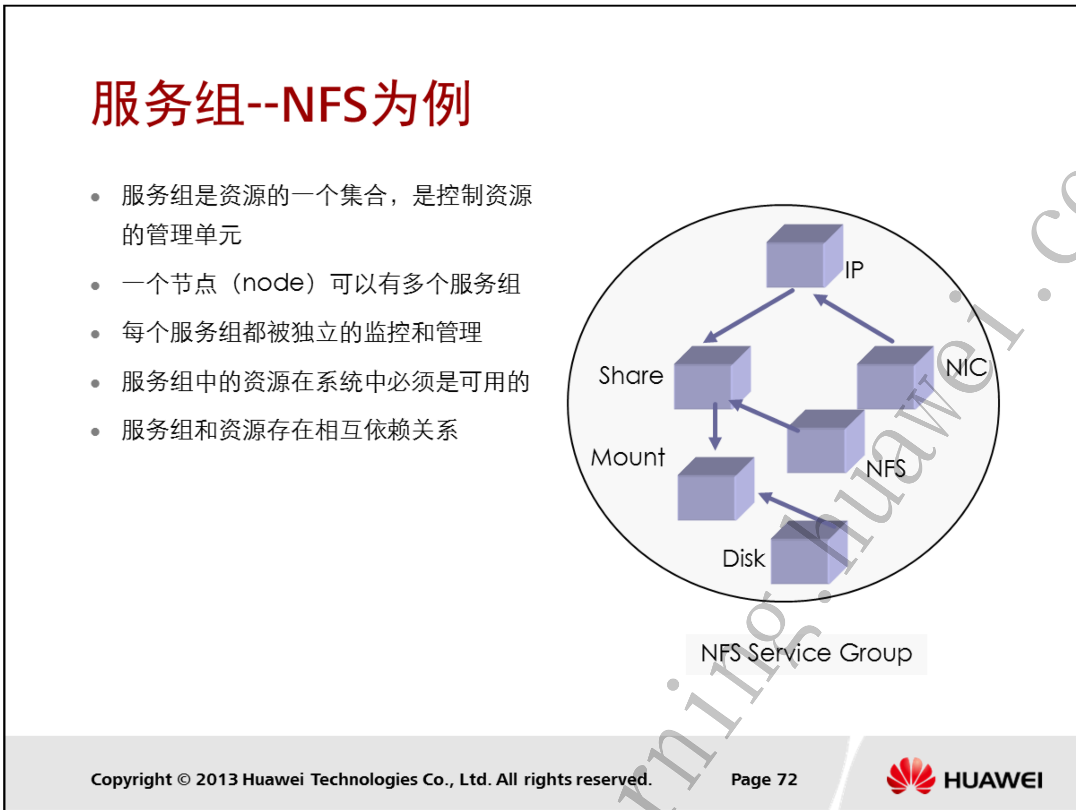
N8500系统提供了全局的命名空间，从而实现了多个节点可以并发访问同一个文件系统。除此之外，N8500系统还提供全局锁机制，保证不同节点访问相同文件系统相同文件并发访问的一致性，通过细粒度的锁设计，保证系统的并发性。

当不同节点访问文件F1的不同区域时，通过该文件的所有者(Owner)进行仲裁，如果没有其他节点访问对应区域，则允许该节点读写数据；如果发现有其他节点访问该区域，则出现冲突，该所有者就会让新请求等待，直到读写该区域的请求完成，然后通知等待该区域的节点竞争锁。通过此机制，实现了集群多节点范围内的一致锁所处理；同时，通过细粒度的区域锁设计，保证高并发。

Agent

- Agent是控制资源的程序
- 每种类型的资源对应一个代理，一个代理控制对应类型的所有资源
- 用户也可以自定义代理





- 服务组状态查看命令：hagrp －state
- 服务组依赖关系查看命令：hagrp －dep
- 服务组包含资源命令： hagrp -resources
- 提示：ha开头的命令都是vcs的命令

服务组的类型

- Failover Service Groups
  - 在某一时间，该service group只能运行在一个系统上（system），在对数据一致性要求高（比如数据库，网络文件系统（NFS））的情况下，使用该组。
- Parallel Service Groups
  - 在某一时间，可以部署在多个系统上 配置比Failover Service Groups复杂很多
- Hybrid Service Groups
  - 是上面两种service groups的组合
- 在N8000中只使用了Failover和Parallel两种

资源（Resource）定义

- 资源是VCS对象，它对应到应用服务的硬件或者软件组件
- 资源是放在服务组里面的定义和管理的。

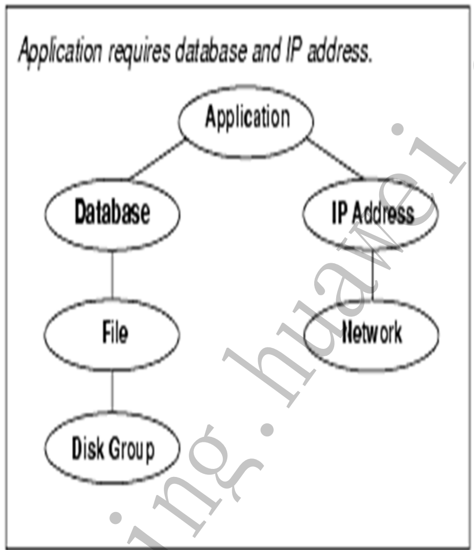
HUAWEI

```
group VIPgroup1 (
    SystemList = { tcl_2 = 1, tcl_1 = 0 }
    AutoStartList = { tcl_2, tcl_1 }
    FailOverPolicy = Load
    AutoStartPolicy = Load
    PreOnline = 1
    Load = 40
    )
    IP VIP1 (Device = bond0 Address = "192.168.34.101 " NetMask = "255.255.255.0 " )
NIC VIPNIC1 (Device = bond0  Mii = 0 )

    requires group Phantomgroup_bond0 online local firm
    VIP1 requires VIPNIC1
```

资源依赖关系

- 资源可以依赖其他资源
- 父资源依赖子资源
- 服务组可以依赖其他服务组
- 资源和服务组之间的依赖关系由规则管理
- 不允许出现循环依赖



HUAWEI

- 查看资源详细信息命令：hares –state以及hares –display
- 查看资源依赖关系：hares –dep

目录

HUAWEI

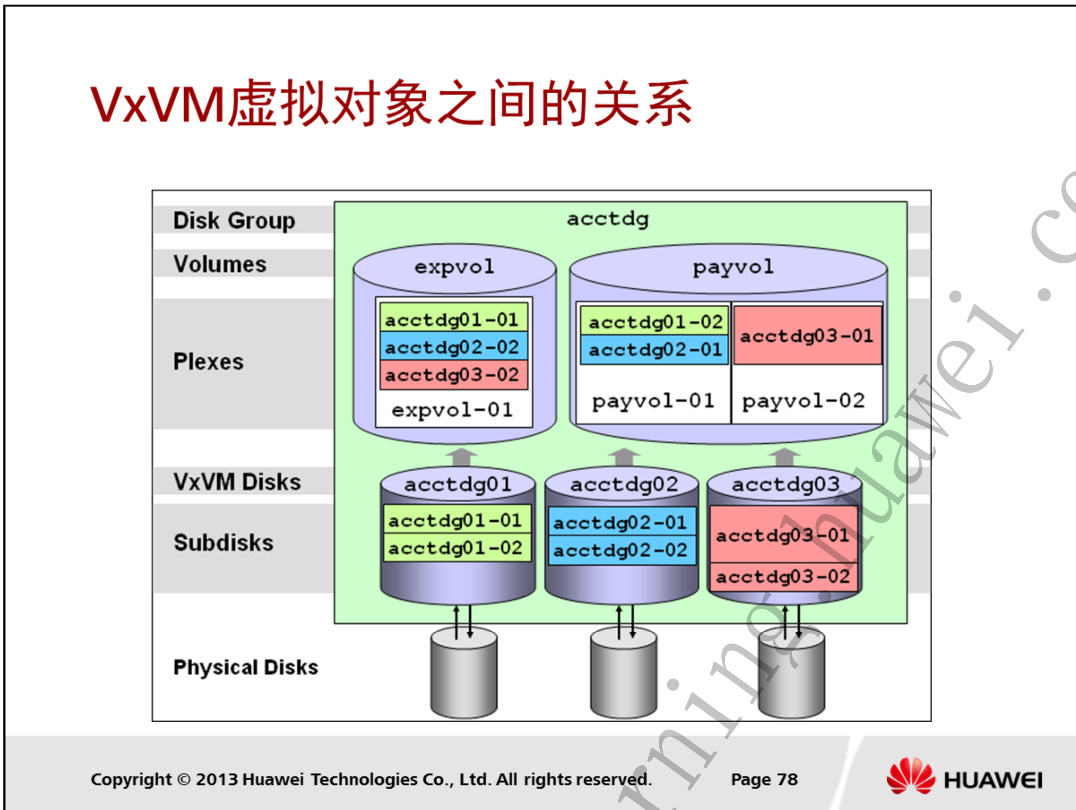
VxVM基本概念

- 物理对象：能被OS看到的物理存储设备
- 虚拟对象：VxVM进行存储设备管理的逻辑对象。VxVM通过虚拟对象和物理设备的映射来访问存储设备
- 虚拟对象有：卷，磁盘组，VM磁盘，子磁盘，Plex
- 卷布局: 卷的布局是指卷与物理存储设备的映射关系，即卷与一个或多个Plex的关联关系。不同的卷布局提供不同级别的存储服务

VM磁盘	一个物理磁盘，分区或者LUN
磁盘组	若干个VM磁盘的组合
子磁盘	VM磁盘上连续的磁盘块
Plex	由一个或多个物理磁盘上的一个或者多个子磁盘组成
卷	由一个或多个Plex组成

 Page 77 HUAWEI

VERITAS Volume Manager位于操作系统和数据管理系统（文件系统、数据库管理系统）之间，利用操作系统提供的接口服务访问磁盘；将连接到主机的物理存储设备屏蔽，为数据管理系统提供了一层称之为卷的逻辑设备；除了提供磁盘管理的基础功能外，还提供了其他一些高级功能。



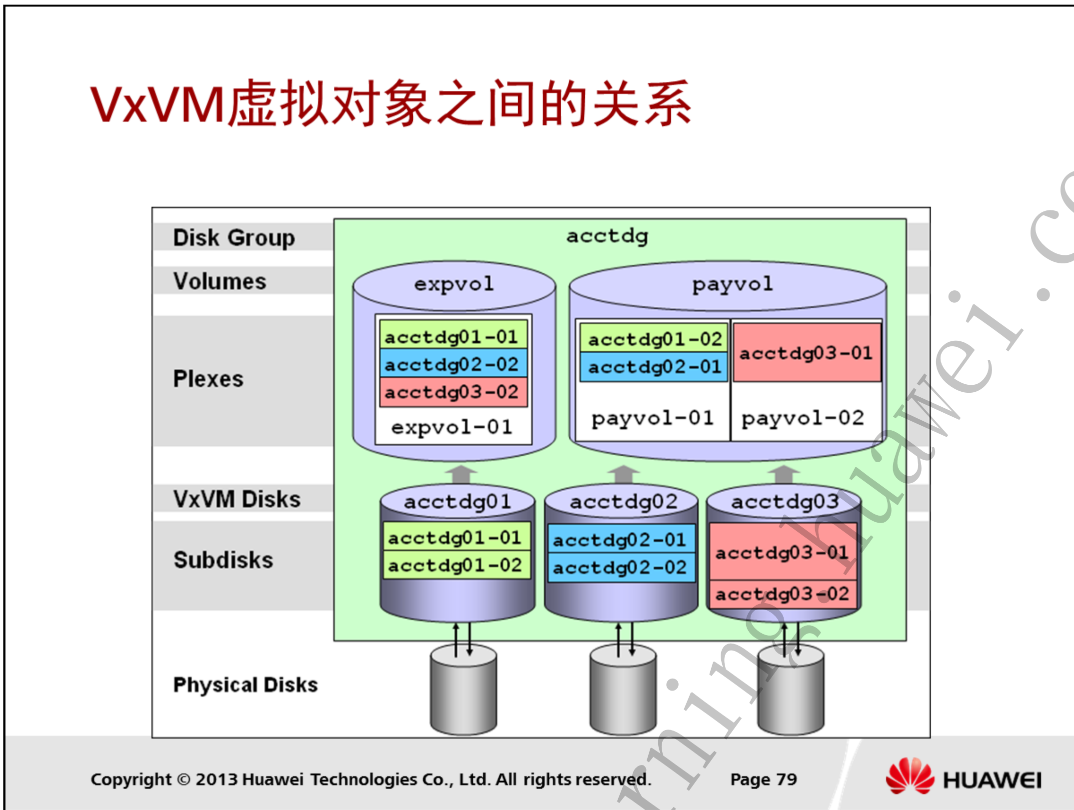
- 磁盘组：是使用公共配置并由VxVM管理的一些磁盘的集合。磁盘组配置是包含相关VxVM对象机器属性和联系的详细信息的记录集。
- VM磁盘：将物理磁盘置于VxVM控制之下时，系统会为该物理磁盘分配一个VM磁盘。VM磁盘受VxVM控制，通常位于某个磁盘组中。每个VM磁盘至少与一个物理磁盘或磁盘分区对应。
- 子磁盘：是一组连续的磁盘块。VxVM使用子磁盘分配磁盘空间，一个VM磁盘可以分成一个或多个子磁盘。一个VM磁盘可以包含多个子磁盘，但子磁盘不能互相重叠或共享VM磁盘的同一部分。
- Plex：VxVM使用子磁盘创建称为plex的虚拟对象。Plex由位于一个或多个物理磁盘上的一个或多个子磁盘组成。
- 卷：是一种虚拟的磁盘设备，就应用程序、数据库和文件系统而言，它与物理磁盘设备很相似，但没有物理磁盘设备的物理局限性。卷由一个或多个plex组成，每个plex都包含卷中所选数据的一个副本。

vt public07 - ENABLED - - ACTIVE - -

v public07_tier1 public07 ENABLED 23622320128 - ACTIVE - -

pl public07_tier1-01 public07_tier1 ENABLED 23622328320 - ACTIVE - -

sd ENGENIO-1_5-01 public07_tier1-01 ENABLED 2851930368 0 - - -



sd ENGENIO-1_10-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

sd ENGENIO-1_6-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_11-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

sd ENGENIO-1_7-01 public07_tier1-01 ENABLED 2851930368 0 - - -

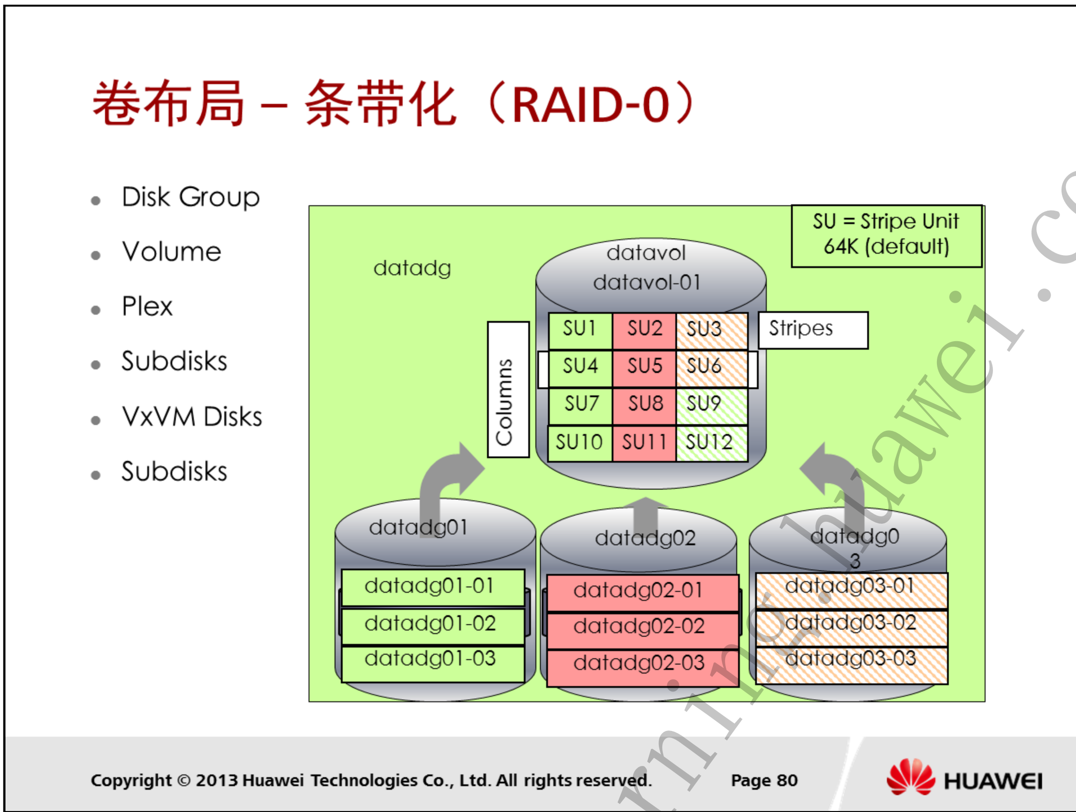
sd ENGENIO-1_12-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

sd ENGENIO-1_8-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_13-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

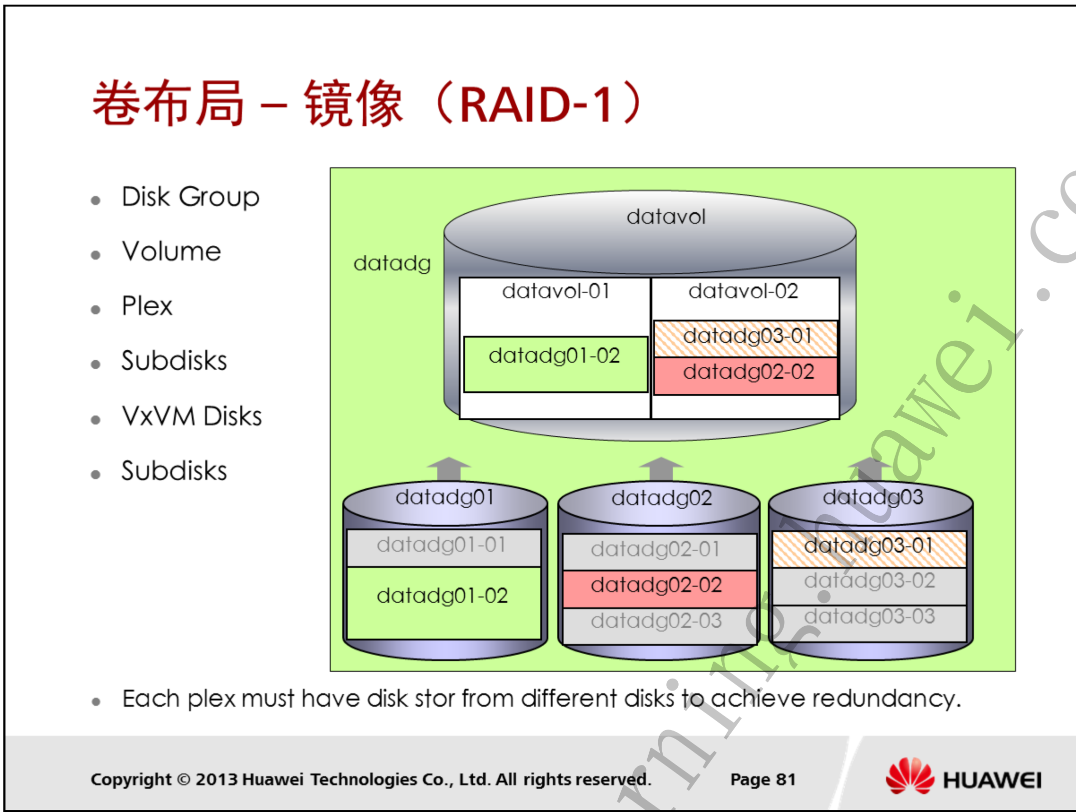
sd ENGENIO-1_9-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_14-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -



条带化（RAID-0）是对数据进行映射，从而使数据交错分布在两个或两个以上的物理磁盘上。一个条带plex包含两个或更多个磁盘，子磁盘可以组合成“列”，每个物理磁盘都被限制在一列中。一个列包含一个或多个子磁盘，并且可以从一个或多个物理磁盘中派生出来。每列的子磁盘的数量和大小可能各不相同。如果需要可以向列中添加更多的子磁盘。

```
vt fs1          -          ENABLED -        -       ACTIVE  -     -
v  fs1_tier1    fs1        ENABLED  608174080 -     ACTIVE  -     -
pl fs1_tier1-01 fs1_tier1  DISABLED 608174080 -     NODEVICE -    -
sd OceanStor0_0-02 fs1_tier1-01 DISABLED 608174080 0  NODEVICE -  -
pl fs1_tier1-02 fs1_tier1  ENABLED  608174080 -     ACTIVE  -     -
sd OceanStor0_1-02 fs1_tier1-02 ENABLED 608174080 0  -     -     -
pl fs1_tier1-04 fs1_tier1  ENABLED  LOGONLY -       ACTIVE  -     -
sd OceanStor0_1-03 fs1_tier1-04 ENABLED 2112  LOG   -     -
```



镜像功能使用多个镜像（plex）复制卷中包含的信息。在物理磁盘发生故障的情况下，故障磁盘上的plex将变得不可用，但系统仍可使用未受影响的镜像继续运行。

镜像的每个plex用到的物理磁盘可以来自不同RAID组或不同存储阵列。

vt public07 - ENABLED - - ACTIVE - -

v public07_tier1 public07 ENABLED 23622320128 - ACTIVE - -

pl public07_tier1-01 public07_tier1 ENABLED 23622328320 - ACTIVE - -

sd ENGENIO-1_5-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_10-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

sd ENGENIO-1_6-01 public07_tier1-01 ENABLED 2851930368 0 - - -

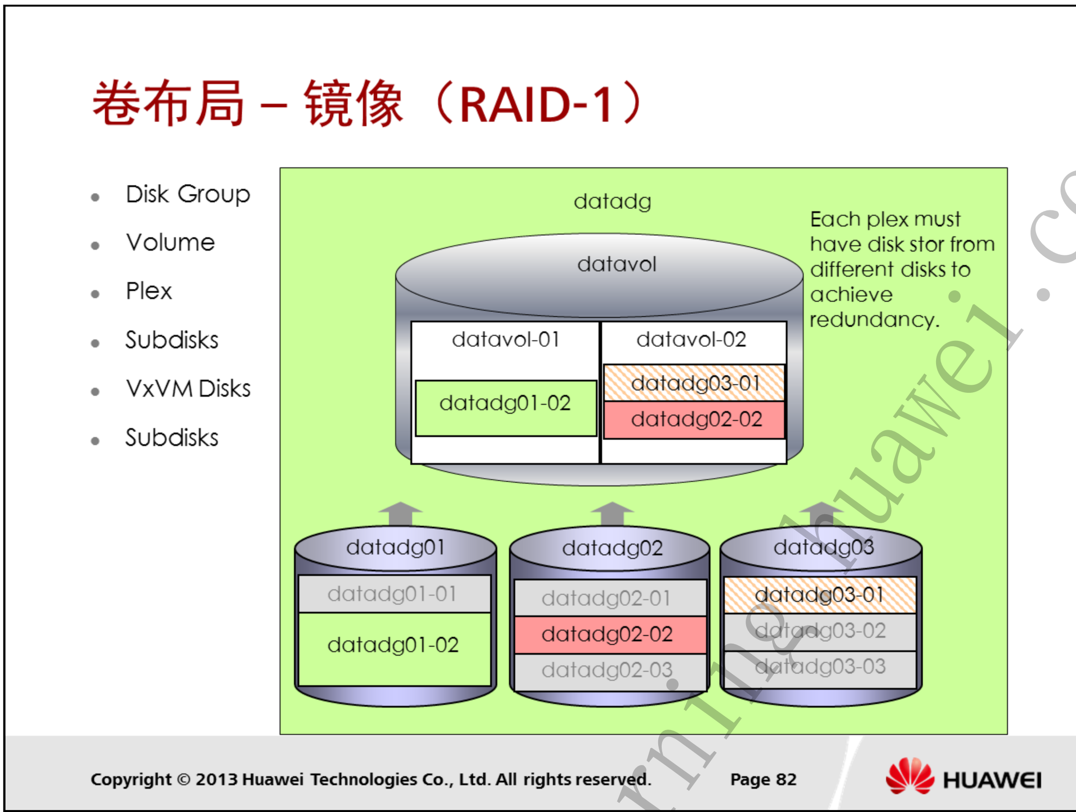
sd ENGENIO-1_11-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

sd ENGENIO-1_7-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_12-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

sd ENGENIO-1_8-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_13-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -



sd ENGENIO-1_9-01 public07_tier1-01 ENABLED 2851930368 0 - - -

sd ENGENIO-1_14-01 public07_tier1-01 ENABLED 1872535296 2851930368 - - -

总结

- 大数据存储架构
- 大数据存储技术
- 大数据分析技术
- 数据安全技术
- NAS基础技术
- 集群NAS技术
- 集群NAS卷管理技术

思考题

- 请描述大数据存储的Scale-out
- 对象存储与传统存储的区别是什么?
- EC技术的冗余原理是什么？如何实现快速恢复
- 集群NAS中如何实现集群节点间信息通信?
- 集群NAS中资源、代理、服务之间的关系是什么?
- 集群NAS中低层虚拟化通过什么实现?

Thank you

www.huawei.com

HC1209203 集群 NAS 系统部署与管理

HC1209203 大数据存储集群NAS系统部署与管理

www.huawei.com

HUAWEI

目标

- 学完本课程后，您将能够:
  - 描述集群NAS硬件组成和软件特性
  - 掌握N8500系统内外组网
  - 掌握N8500软硬件安装
  - 掌握NFS和CIFS配置部署
  - 了解高级特性原理与配置
  - 了解集群NAS应用方案

目录

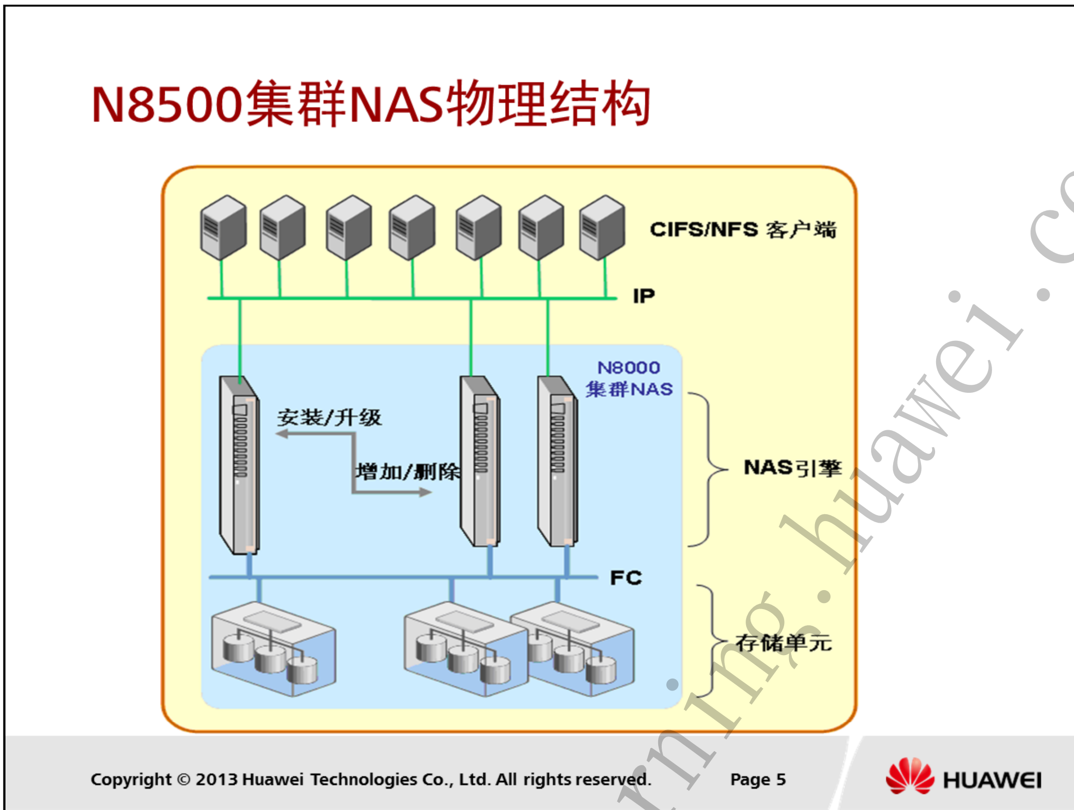


- **高性能**
  - 性能领先：N8500在SPEC sfs2008基准评测中性能业界第一，其中NFS协议下性能达到3064602 OPS
  - 集群架构：多节点全Active集群技术，集群中所有节点可协同工作，实现对数据的并发访问
  - 全局共享存储空间：所有节点可以同时访问相同的存储空间；通过DNS实现负载均衡，将负载平均分配至集群节点上，提升节点资源利用率
  - 高可靠架构设计：所有部件均为冗余设计，无单点故障；数据保险箱技术、文件系统镜像等软件技术进一步提高系统可靠性
- **灵活扩展**
  - 引擎节点可扩展：引擎节点可最大扩展至24节点，系统性能随引擎节点增加线性增长
  - 存储容量可扩展：存储容量可扩展到高达15PB
- **高效数据管理**
  - 动态分级存储功能：基于文件访问频率、时间策略在不同存储介质之间实现双向动态迁移，对应用透明，降低企业投资成本
  - 动态数据缓存：通过使用闪存驱动器来扩展缓存并配合动态数据缓存技术，提升系统性能



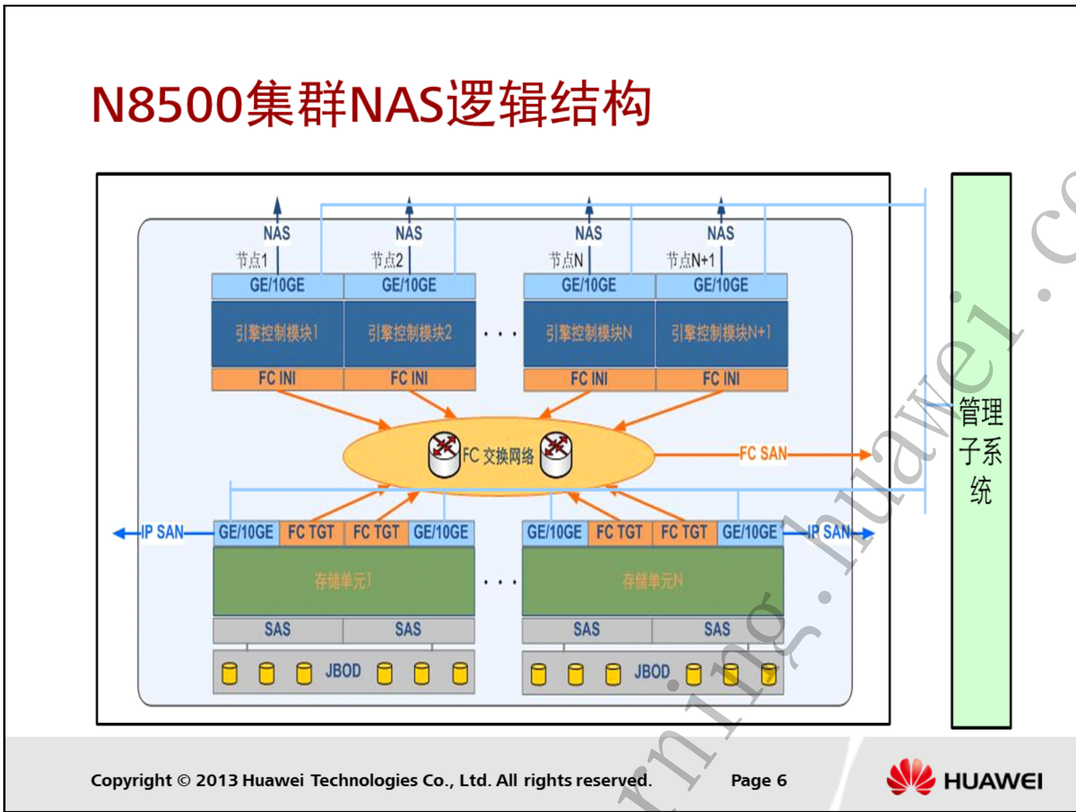
- 文件系统镜像功能：同一存储系统内可设置文件系统镜像，保存多份数据，提高在线数据的可靠性
- 软配额/硬配额管理：实现灵活的用户空间分配和使用
- 域管理：支持AD/NIS/LDAP
- 多种数据保护技术：提供快照、拷贝、远程复制等多种增值特性

- **统一存储**
  - 协议统一：包含SAN和NAS存储协议的统一存储，同一存储系统支持结构化和非结构化数据存储。支持iSCSI、FC、FCoE、NFS、CIFS、HTTP、FTP等多种存储网络协议
  - 管理统一：通过简洁的图形化界面，提供文件、块数据业务的统一管理。向导式的配置方式，协助用户轻松完成各项配置
  - 设备监控：图形化管理界面提供实时/历史性能统计功能，支持声、光、邮件等多种告警方式，随时掌控设备运行情况

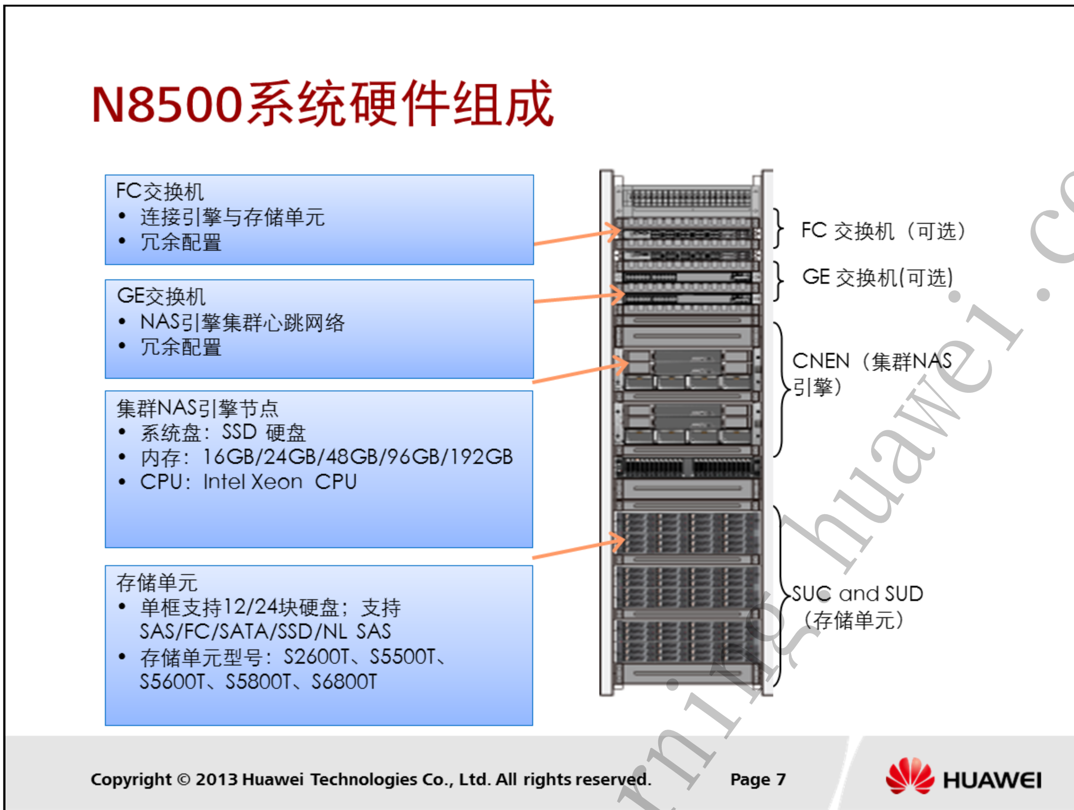


在N8500集群NAS中，集群节点间的通讯通过专用的心跳网络来实现，引擎节点可以在线扩展，性能随着节点扩展线性增加；引擎与存储单元之间通过光纤交换设备连接，为集群NAS提供灵活扩展的存储空间。

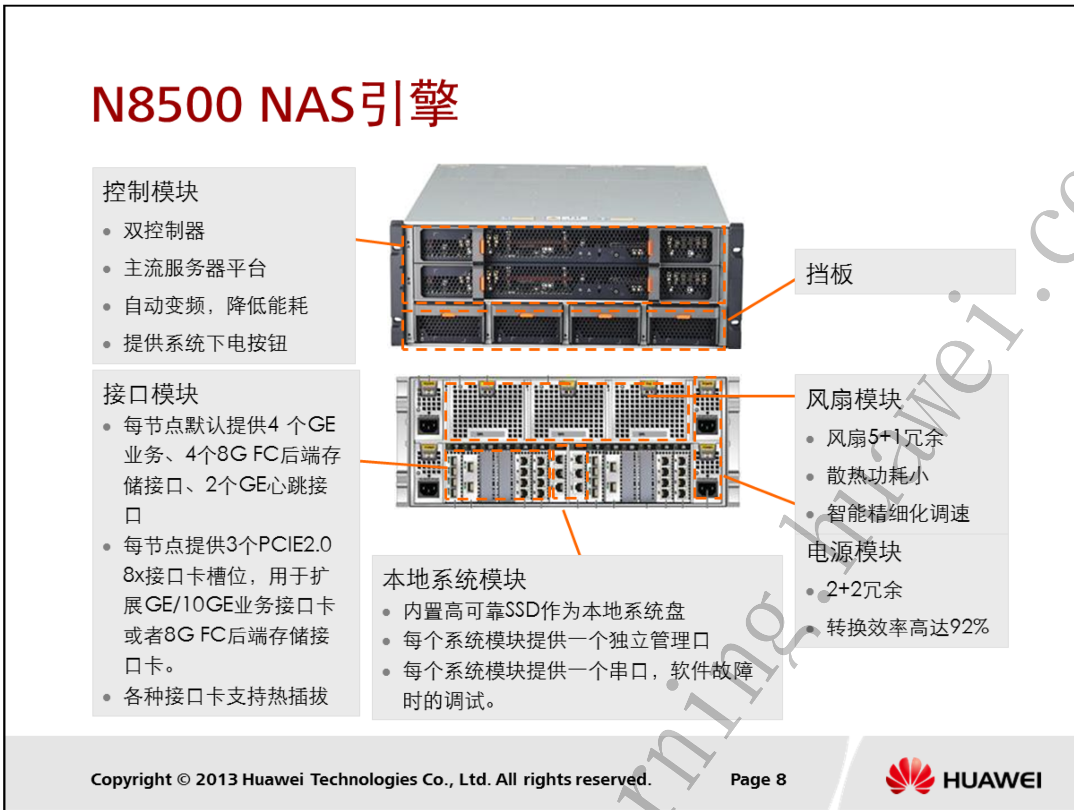
FC SAN作为N85000引擎与存储单元间的内部组网；引擎通过IP网络向外提供NFS和CIFS文件共享服务。



- 集群NAS引擎节点间通过1GE连接心跳网络，完成集群控制信息通信；
- 每个节点对外提供1GE或10GE，提供NAS存储服务和文件业务数据传输；
- 每个集群节点与后端存储单元通过FC网络连接，组成SAN网络，引擎为发起端，存储为目标段；
- 存储控制器与硬盘（硬盘框）之间通过SAS网络连接，组成冗余网络。
- 存储单元其余iSCSI或FC主机端口可以直接连接业务服务器，提供SAN存储服务。



- N8500系统硬件主要组成：
  - 集群NAS引擎机框，每个集群NAS引擎机框包括2个集群NAS引擎节点（CNEN）
  - 存储单元，包括存储单元控制框（SUC）和存储单元硬盘框（SUD）。
  - 以太网交换机 （可选）
  - 光纤交换机 （可选）



- 中间的管理模块 第一个口是管理网口，第二个未使用，第三个是串口


N8500存储单元-- S5600T/S5800T/S6800T
控制模块
• 双控制器
• 主流服务器平台
• 自动变频，降低能耗
• 提供系统下电按钮
接口模块
• 12个接口卡槽位
• 最大支持1440盘
• 接口卡支持热插拔
• 接口类型丰富：8G FC、4G FC、6G SAS、GE、10GE
BBU模块
• 2+2冗余
• 支持直流/交流电源掉电保护
风扇模块
• 风扇5+1冗余
• 散热功耗小
• 智能精细化调速
电源模块
• 2+2冗余
• 转换效率高达92%
管理模块
• 1+1冗余，支持热插拔
• 内置高可靠SSD作为系统保险箱
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 9
HUAWEI

N8500存储单元—S5500T控制框



2.5inch硬盘插框

- 24×2.5inch硬盘；支持主流SAS、SATA、SSD盘；前四个盘为保险箱盘，2+2冗余；每个硬盘可独立上下电

3.5inch硬盘插框

电源模块

- 1+1冗余
- 转换效率最高达92%

盘控一体控制模块

- 主流桌面CPU平台
- 提供管理网口和串口
- 板载2个6Gb SAS接口
- 支持1个iSCSI接口卡
- 板载4个8G FC接口

风扇/BBU模块

- 风扇3+1冗余
- 智能精细化调速
- 内置BBU，支持掉电数据保护

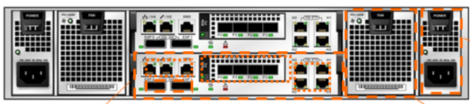
HUAWEI

N8500存储单元--S2600T控制框



3.5inch硬盘插框

- 12×3.5inch硬盘
- 支持主流SAS、SATA、SSD盘
- 前四个盘为保险箱盘，2+2冗余



电源模块

- 1+1冗余
- 转换效率最高达92%

盘控一体控制模块

- 提供管理网口和串口
- 板载2个4*6Gb SAS接口
- 支持1个扩展I/O模块
- 板载6个GE接口

风扇/BBU模块

- 风扇智能精细化调速
- 内置BBU，支持掉电数据保护

HUAWEI

N8500交换组件—FC/GE交换

N8500 FC交换单元SNS交换机



管理端口

8Gb 端口，最大支持20口
(内部连接NAS引擎和存储)

堆叠端口，可由
10Gb升级至
20Gb

N8500 GE交换单元S5324交换机



电源接口
直流/交流

GE端口，用于两节点以上
NAS引擎心跳组网

管理端口

N8500系统关键指示灯

- IT设备默认绿色指示灯代表正常，红色代表故障或异常
- 以下重点描述几个有用指示灯状态：

检查项	指示灯名称	状态和说明
电源模块	运行/告警指示灯	红色，闪：输入欠压（包括交流断电）或过压。
NAS业务模块	电源指示灯	绿色，闪：系统启动阶段。
系统插框	电源指示灯	绿色，闪：机框正在上电或下电。
系统管理模块	管理网口active指示灯	橙色，闪：正在传输数据。
接口模块	NAS业务10GE接口卡端口指示灯	蓝色，亮：数据传输速率为10Gbit/s。
	后端存储FC接口卡端口link/speed指示灯	蓝色，亮：数据传输速率为8Gbit/s。 绿色，亮：数据传输速率为4Gbit/s或2Gbit/s。

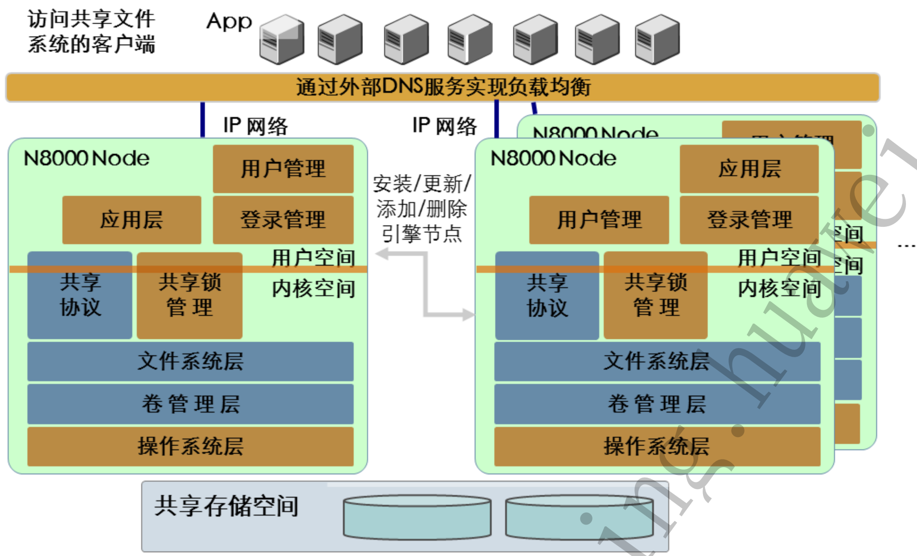
N8500 硬件特性

硬件特性	
系统架构	多节点全Active集群架构
节点数	2～24
缓存/每节点	基础版：16GB；标准版：24GB
	企业版：48GB；增强版：96GB/192GB
GE端口/每节点	标配4 x GE，可扩展至16 x GE
10GE端口/每节点	最大支持6 x 10GE
支持的硬盘类型	SSD、SAS、SATA、NL-SAS
物理存储容量	可动态扩展，最大15PB

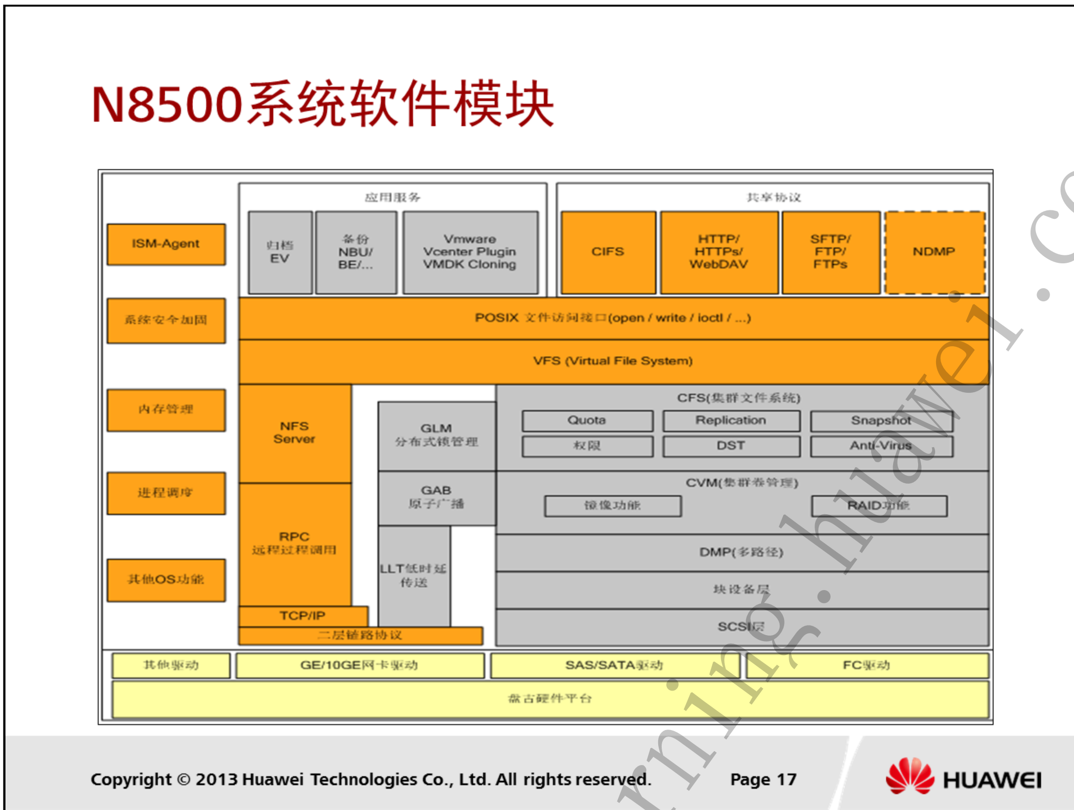
目录

  HUAWEI

集群NAS N8000软件架构



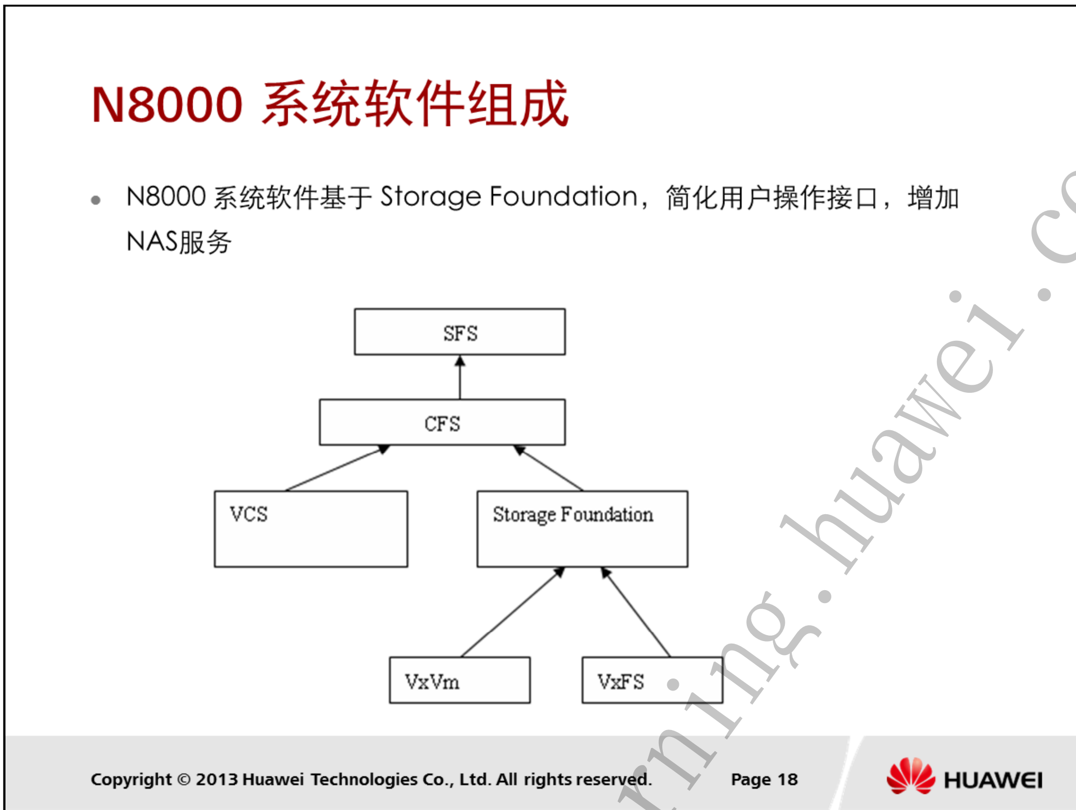
 Page 16 HUAWEI



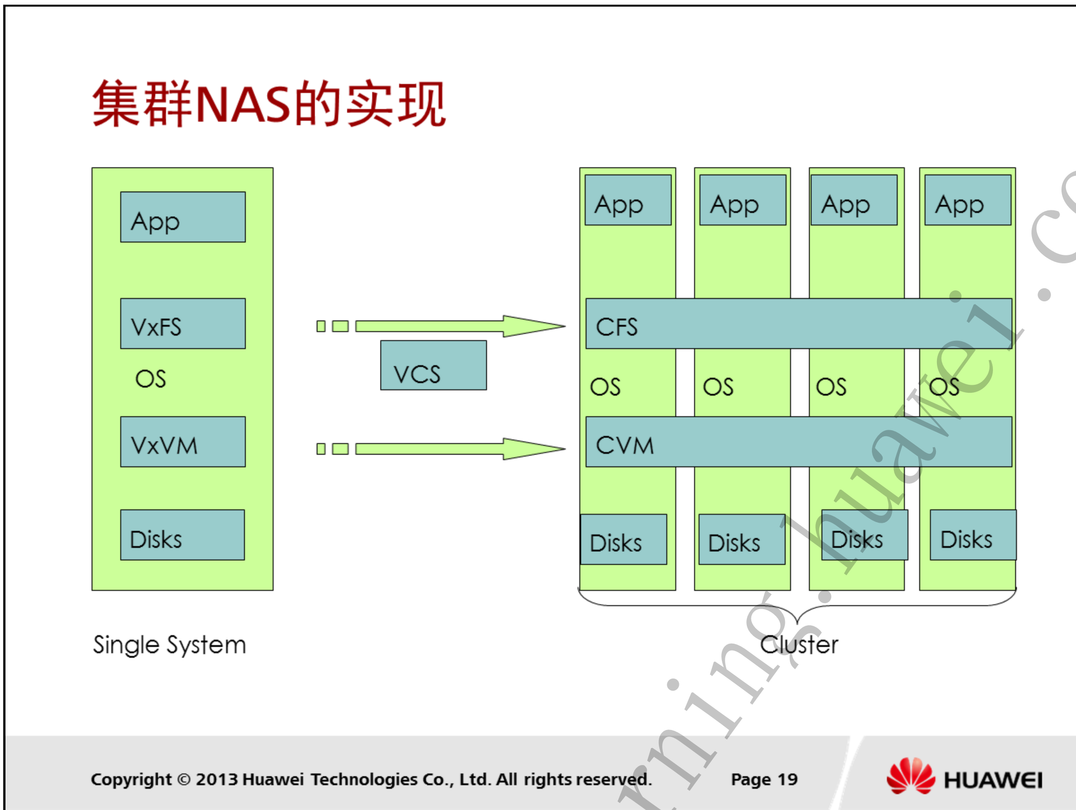
N8500的集群NAS软件基于商用的CFS开发，通过在专用64位存储系统上安装CFS软件，并配置NFS和samba等服务以对外提供NAS访问接口。

N8500集群NAS软件的核心部件CFS提供了大量的增值功能，如快照、动态分级存储等功能。FileStore软件及应用架构示意图如图所示。FileStore软件由核心层的CFS、CVM软件和用户层的各类NAS服务应用组成。FileStore软件部署于集群NAS引擎节点（Node）上，集群NAS引擎节点可通过集群方式扩展，提供全Active的集群架构。所有节点可访问同一个共享存储空间，对外可通过以太网向上层应用提供NAS业务。

N8500集群NAS存储系统提供GUI和CLI两种管理方式。GUI支持NAS引擎和存储单元统一管理，CLI仅支持NAS引擎和存储单元独立管理。另外有些高级功能及部分系统配置功能仅适用CLI支持，面向高级管理维护人员。



- N8500系统软件基于Veritas Storage Foundation和Veritas Cluster Server（VCS），简化用户操作接口，增加NAS服务。
- Storage Foundation是VxVM（卷管理）和VxFS（文件系统）组合构成的，都属于存储，VCS是集群软件，管理SFS。
- CVM和CFS跟VM和FS有联系的，FS在VCS的管理下成为集群文件系统CFS，VM就是VCS管理的VM，集群卷管理。
- VxVM：卷管理组件，负责存储管理功能。
- VxFS：文件系统组件，负责文件管理。
- VCS：集群管理组件，负责集群资源的管理。


集群NAS的实现
App
VxFS
OS
VxVM
Disks
Single System
VCS
App
App
App
App
CFS
OS
OS
OS
OS
CVM
Disks
Disks
Disks
Disks
Cluster
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 19
HUAWEI

N8500软件特性

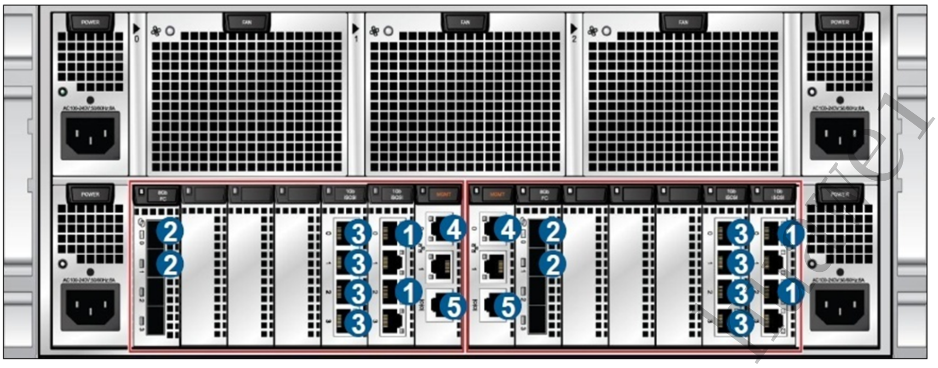
软件特性	
RAID支持	0、1、3、5、6、10、50
文件系统大小	可动态扩展，最大256TB
支持快照数量	512（每个文件系统）
存储协议	FC、FCoE、iSCSI、NFS、CIFS、FTP、HTTP
访问控制	NIS、Microsoft Active Directory、LDAP
其他协议	NDMP、SNMP、NTP
增值功能	DST（动态分级存储）、Snapshot、Mirror、Quota、Replication、SmartCache（动态数据缓存）、HyperThin（自动精简配置）、HyperImage（快照）、HyperCopy（LUN拷贝）、HyperClone（分裂镜像）、HyperMirror（同步/异步远程复制）
配额	支持NFS/CIFS的软/硬配额管理
管理方式	GUI、CLI
兼容性	Windows、Linux、Mac OS、Solaris、AIX、HP-UX等主流操作系统

HUAWEI

目录

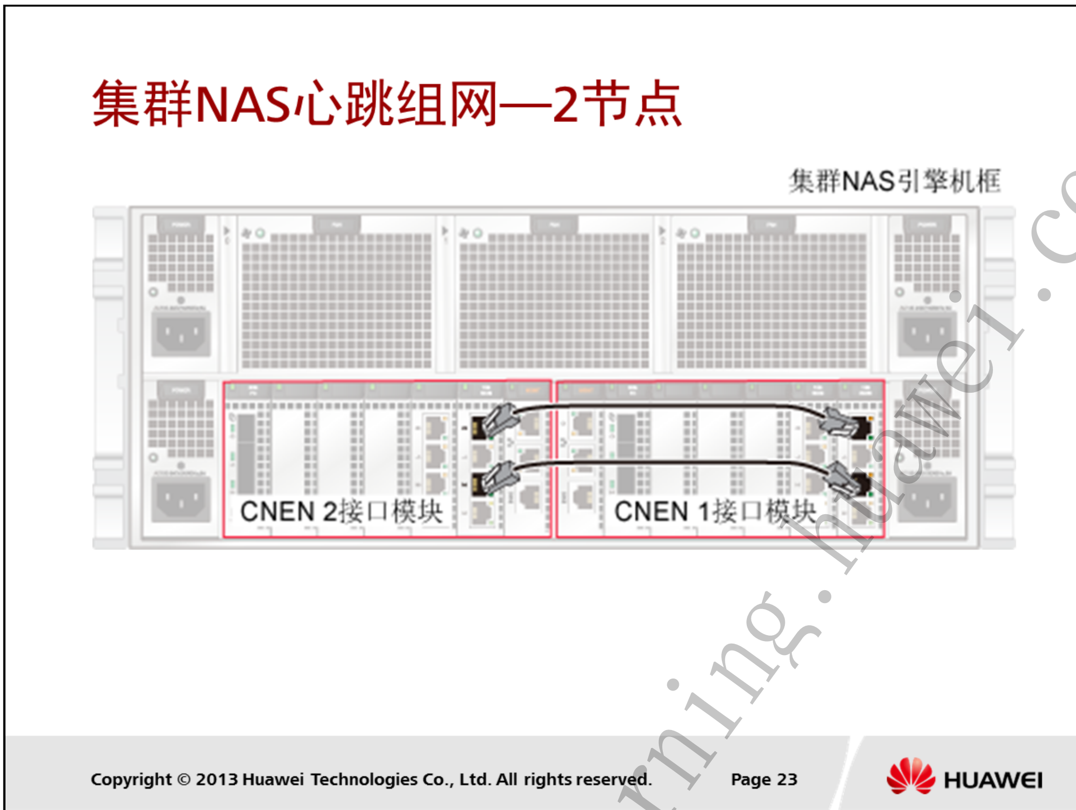
HUAWEI

N8500模块接口

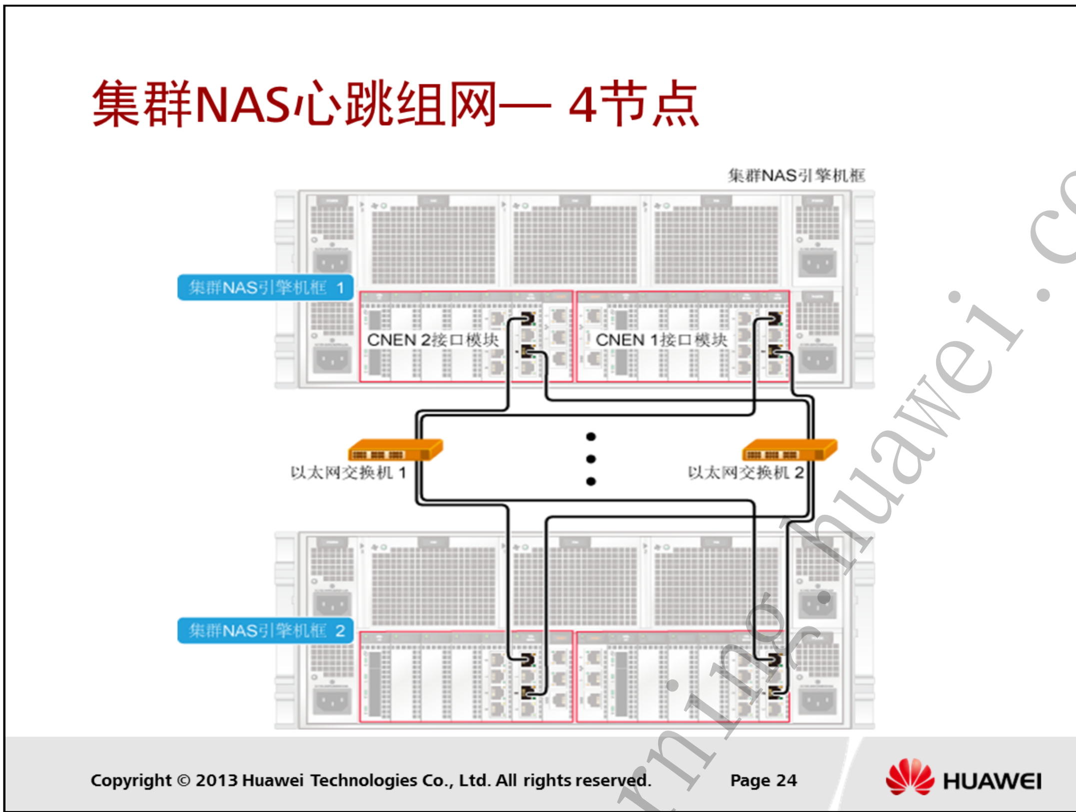


 Page 22 HUAWEI

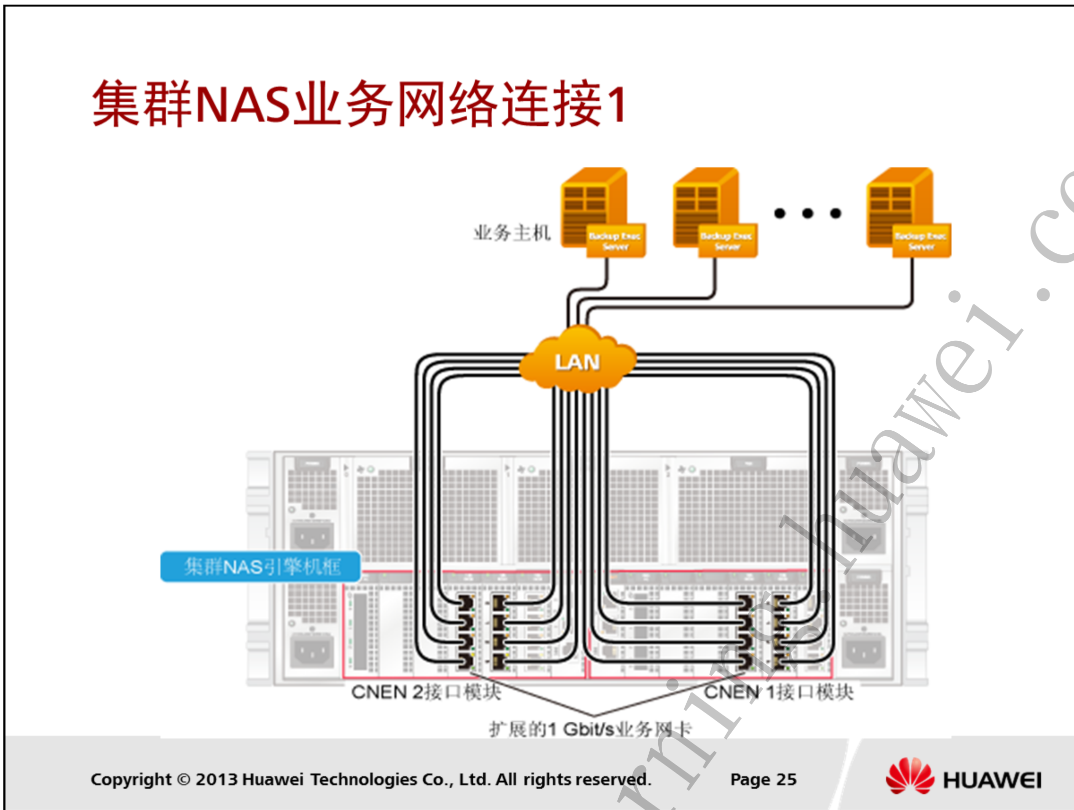
- 1代表心跳网口，采用1Gb iSCSI 组建心跳网络，每个集群NAS引擎节点提供编号为P0和P2的2个网口用于心跳连接，该网口只用于心跳连接，此接口卡的其他两个网口不使用。
- 2代表存储接口，采用8Gb FC，通常通过交换机交换机连接存储，每个集群NAS引擎节点提供2个，建议按端口顺序使用。
- 3代表业务网口，采用1Gb iSCSI、10Gb iSCSI，建议按网口顺序使用，两个集群NAS引擎节点必须提供相同数量、类型、位置的网口。
- 4代表管理网口 MGMT，管理N8500 业务系统，需连接所有集群NAS引擎节点管理网口到维护终端所在的网络。
- 5代表管理串口，通过超级终端进入CLI界面，波特率是115200，其他默认。



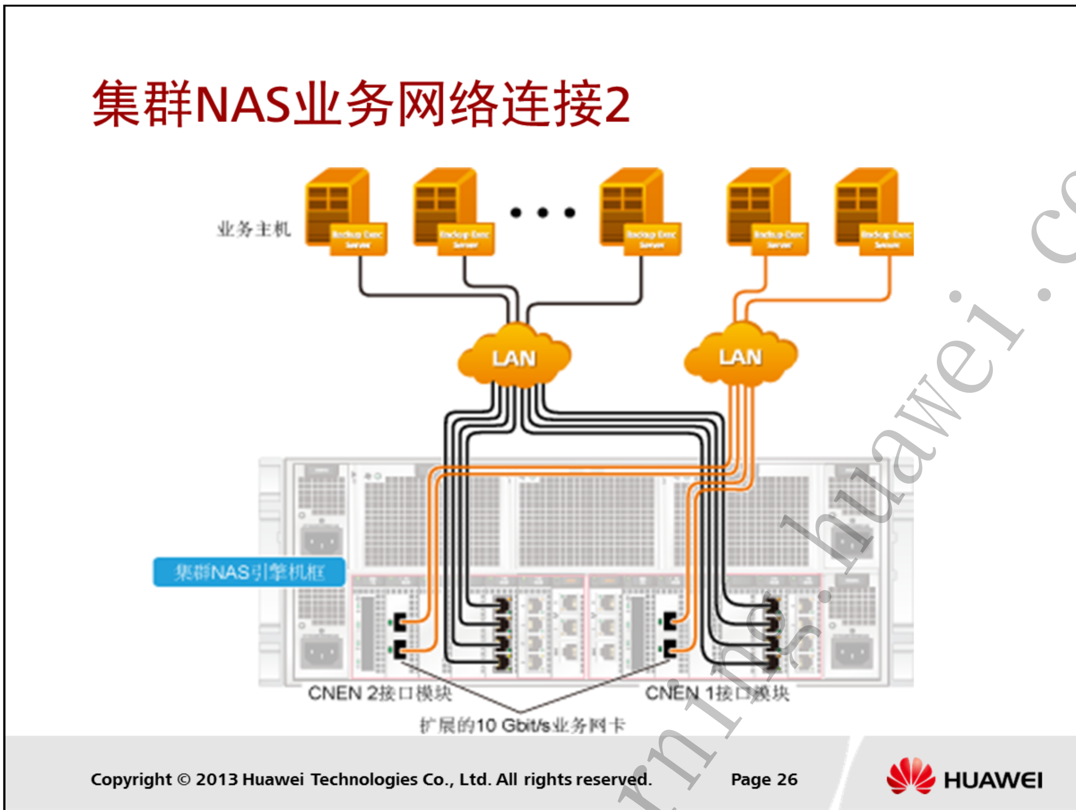
集群NAS引擎配置2个CNEN时，心跳网络可采用直连的方式。每个集群NAS引擎节点提供编号为P0和P2的2个网口用于心跳连接，该网口只用于心跳连接，此接口卡的其他两个网口不使用。



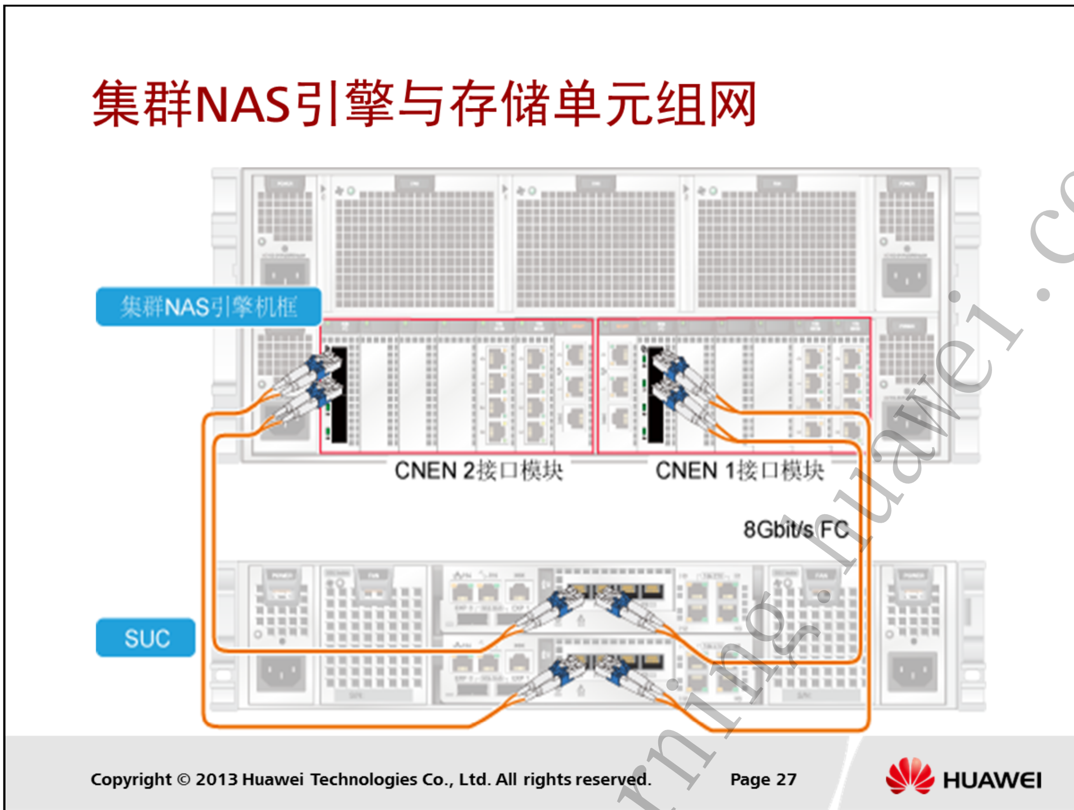
集群NAS引擎配置3个～24个CNEN时，必须为心跳网络配置2台以太网交换机，组建多台CNEN的心跳网络。



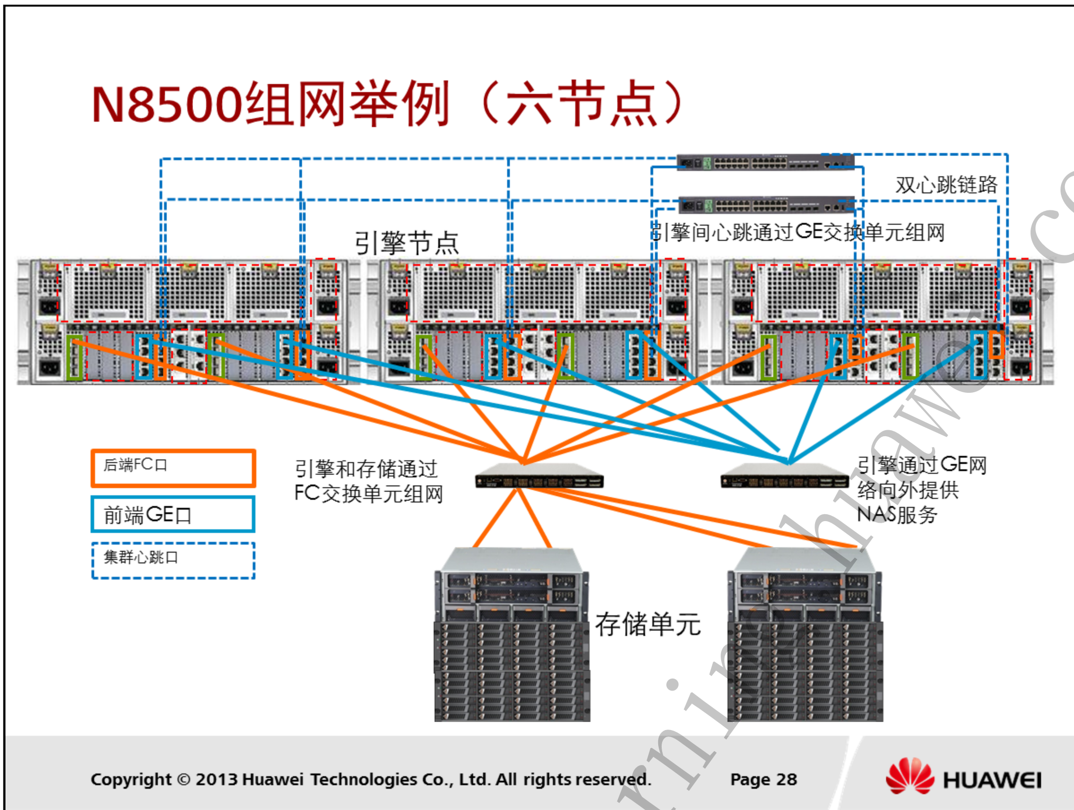
- 单个CNEN缺省配置的业务网口数量为4个。
- CNEN可以与业务主机直连，也可以连接至业务网络中的交换机：
- 每个CNEN可以提供4个1Gbit/s业务网口用于连接至业务网络。
- 每个CNEN可扩展1张～3张1Gbit/s网卡，最多提供16个1Gbit/s业务网口用于连接至业务网络。



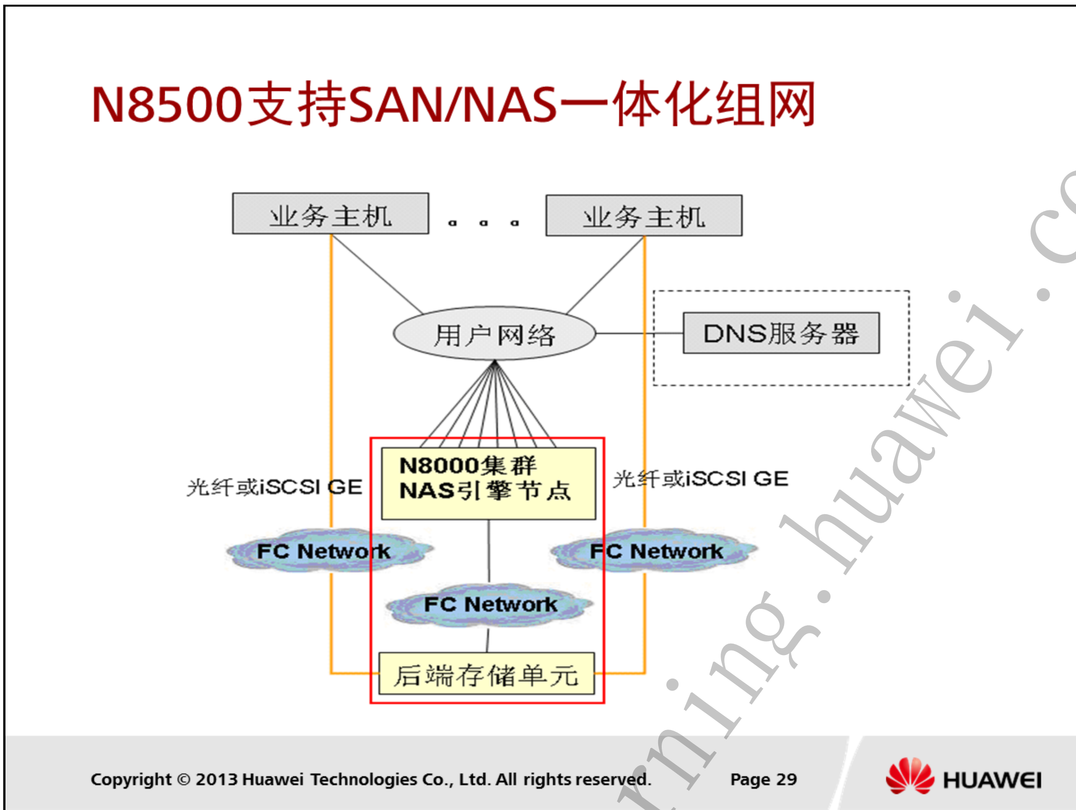
- 每个CNEN可扩展1张～4张10Gbit/s网卡，最多提供8个10Gbit/s业务网口用于连接至业务网络。
- 当需要扩展4张10Gbit/s网卡时，其中一张10Gbit/s网卡将替换缺省配置的1Gbit/s网卡。



- 直连组网方式
  - CNEN与SUC可进行直连，或通过光纤交换机进行组网。直连组网时，支持的配置为： 支持2个～4个CNEN。
  - 支持1台～4台SUC。
  - 直连组网方式下，不支持在线扩展CNEN、SUC。
- 交换组网方式
  - CNEN与SUC通过光纤交换机组网时，支持的配置为： 支持2个～24个CNEN。
  - 支持1台～18台SUC。
  - 交换组网方式下支持在线扩展CNEN、SUC。
- 存储单元控制框（SUC）与硬盘框（SUD）级联组网如下：
  - 1台SUD与1台SUC级联时，使用SUC两个控制器上的“EXP 0”端口分别与SUD两个级联模块上的“PRI”端口连接，组建一个级联环路。
  - 2台SUD与1台SUC级联时，使用SUC两个控制器上的“EXP 0”端口分别与1台SUD两个级联模块上的“PRI”端口连接，组建第一个级联环路，使用SUC两个控制器上的“EXP 1”端口分别与另1台SUD两个级联模块上的“PRI”端口连接，组建第二个级联环路。

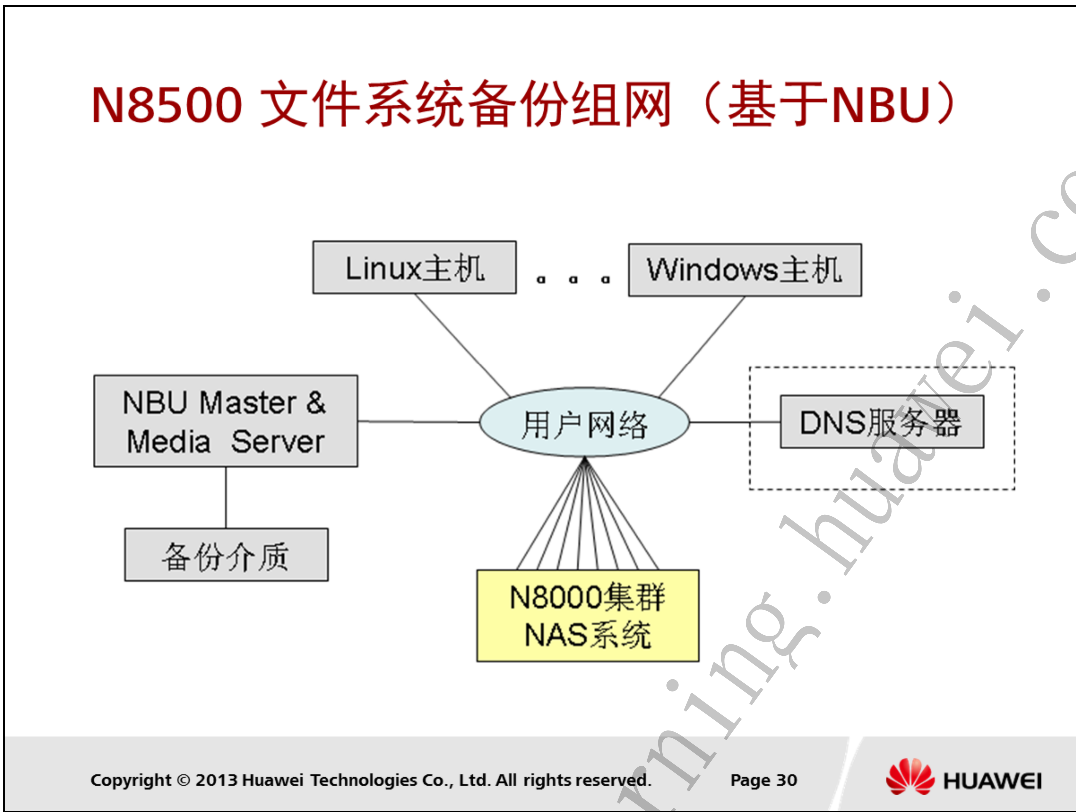


N8500组网举例（六节点）
双心跳链路
引擎间心跳通过GE交换单元组网
引擎节点
后端FC口
前端GE口
集群心跳口
引擎和存储通过
FC交换单元组网
引擎通过GE网
络向外提供
NAS服务
存储单元
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 28
HUAWEI

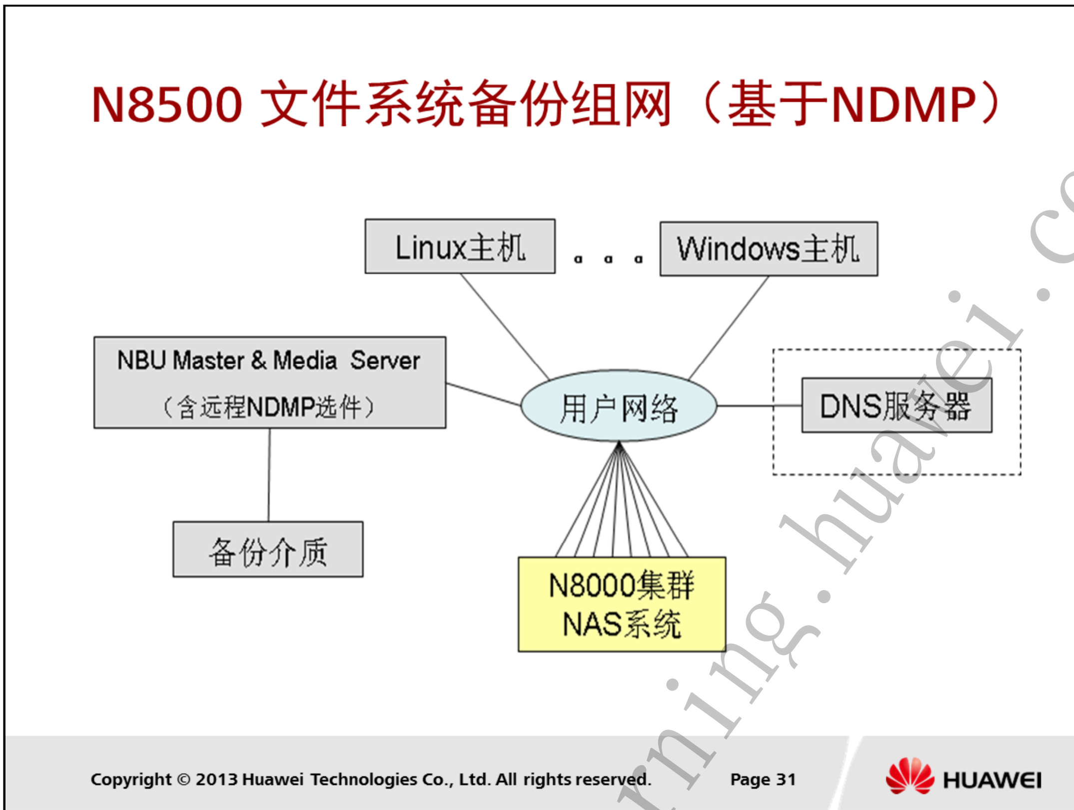


N8500系统的NAS、SAN一体化体现为从后端单元直接提供FC或iSCSI接口，提供FC-SAN和IP-SAN的组网方式供用户业务使用，使用S2600T、S5500T、S5600T、S5800T或S6800T时，由于后端单元可提供较多的接口，在提供N8500集群节点使用之余，其余接口可作为SAN接口使用。

在N8500系统中，仅使用了存储单元的部分端口，此外，在引擎节点数量比较少或使用交换形态的情况下，可以根据需要配置10Gbps iSCSI扣卡，从而为用户提供万兆IP SAN；若用户有SAN的需求，则可以在阵列上将未使用的磁盘空间（以LUN为单位）通过空闲的FC端口或iSCSI/10Gbps iSCSI端口提供给用户使用，达到NAS/SAN一体化的效果，在这种组网下，用户的具体网络环境可能为无域的环境、AD域环境、NIS域环境或LDAP域环境等。



使用NBU客户端进行备份时，需要环境中包含NBU服务器(包括Master Server和Media Server)，如果使用DNS，则要求DNS服务器能对NBU服务器进行正确解析，网络连接好后在N8500系统上配置NBU服务器相关的信息，然后在NBU Master Server上进行备份策略等配置后，备份即可按照设置的策略正常进行，数据流从N8500引擎经Media Server后备份到后端介质，无需再进行干预。



使用NDMP协议进行备份时，当前版本支持通过LAN进行备份的方式，即备份数据通过IP网络进行传输，备份采用的协议为标准的NDMP协议。在采用基于IP网络的NDMP备份时，同样需要有备份服务器（如NBU），由备份服务器连接备份介质（磁带库设备），根据备份软件的不同，可能需要相应的选件支持，如使用NBU时需要备份服务器上安装有远程NDMP选件，在进行备份时，NDMP控制信息由备份服务器下发到N8500系统，N8500系统根据控制命令进行备份，将备份数据通过IP网络传送到备份服务器的相关模块（如NBU的Media Server）再由该模块写入备份介质。

目录



- 管理员功能：管理员具有添加用户、删除用户、修改用户的密码等功能。
- 集群管理：用于管理集群，其中包括支持查看已有结点的功能；节点的动态添加、删除；节点自动安装等。
- 网络配置：网络配置用于管理网络配置，包含的功能有IP的添加、删除和修改；节点配置的修改功能：online，offline；路由的添加、删除；DNS配置；NIS、LDAP的身份认证方式的启用或停用等。
- 历史命令查询：不同用户查看的历史命令权限不同。
- NFS共享：可提供文件系统的只读、读写等方式NFS共享。
- CIFS共享：可提供文件系统的只读、读写等方式CIFS共享。有normal和homedir两种方式，其区别主要在于homedir共享可对共享用户设定用户配额。
- 网管功能：网管功能提供事件的导出，日志查看及导出，SNMP协议支持及邮件告警。
- 存储管理：可进行存储池的添加、删除、重命名；向已有存储池中添加和删除磁盘；文件系统的容量扩展功能（包括增加和缩小）；可创建快照，快照资源 的配额可设定；可创建镜像，镜像方式分为mirrored、mirrored-stripe、simple、striped、striped-mirror等；镜像数量可设置。
- 系统管理：系统管理可进行日期、时间和时区的更改，以及配置文件的本地或远程的导出和导入。
- 技术支持：该功能用于系统信息搜集，便于技术支持定位问题。



- 升级及补丁管理：支持系统在线升级及打补丁。
- 备份功能：支持NBU Client内置备份和NDMP备份协议。
- FTP功能：支持将N8000设置为FTP服务器，对外提供FTP服务。
- 远程复制功能：支持将本地N8000上的文件系统拷贝到远程N8000上。

N8000部署前IP地址规划

- 引擎物理IP规划：

预留物理IP地址数目 = 引擎节点数目 × 单个引擎节点业务网口数目

- 集群NAS引擎虚拟IP规划：

预留虚拟IP地址数目 = 引擎节点数目 × 单个引擎节点业务网口数目 + 1

- 存储单元控制框管理IP规划：

预留IP地址数目 = 存储单元控制框数目 × 2

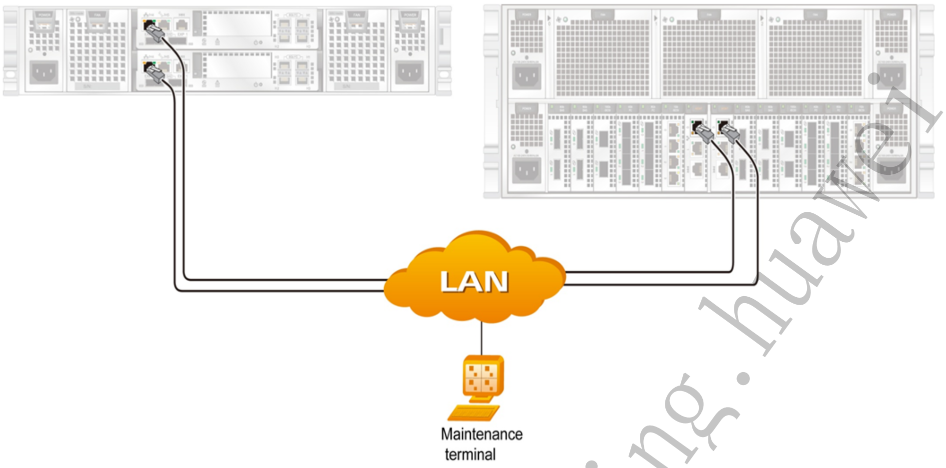
- 以上3类IP地址段不能重叠，同时业务后续扩容，预留一定IP地址空间

- 规划原则：
  - 集群NAS引擎物理IP地址段必须为连续的IP地址段，数量等于集群NAS引擎节点用于组建业务网络的网口
  - （考虑扩展）数量加上集群NAS引擎所有管理网口数量。
  - 集群NAS引擎虚拟IP地址段必须为连续的IP地址段，且数量等于集群NAS引擎物理IP地址的数量。
  - 每套N8500只需规划一个引擎Console IP地址（虚拟IP地址），用于管理N8500。
  - 存在多台存储单元控制框的情况下，需对每一台存储单元控制框规划不同的管理网口IP地址。
  - 集群NAS引擎节点物理IP地址、集群NAS引擎节点虚拟IP地址和集群NAS引擎Console IP地址建议在同一网段内；存储单元控制框管理网口IP地址必须与集群NAS引擎Console IP地址在同一网段内。
  - 存储单元硬盘框不需配置IP地址，因此在网络IP地址规划时不用考虑。



- 系统正常上电顺序：
  - 外部电源通电（此时光纤交换机和存储单元硬盘框即上电） → 以太网交换机依次上电（打开电源开关） → 存储单元控制框（按存储单元控制框前面板的电源按钮） → 集群NAS引擎机框依次上电。
  - 设备上电过程需要大约5分钟 ～ 7分钟，集群NAS引擎机框电源指示灯不再持续闪烁时，表明启动完成。
- 系统正常下电顺序：
  - 停止业务 → 集群NAS引擎机框依次下电（长按集群NAS引擎机框前面板的电源按钮5秒） → 存储单元控制框（长按存储单元控制框前面板的电源按钮5秒） → 存储单元硬盘框依次下电（断开电源线） → 光纤交换机依次下电（断开电源线） → 以太网交换机依次下电（关闭电源开关） →外部电源断电。

N8500管理组网



HUAWEI

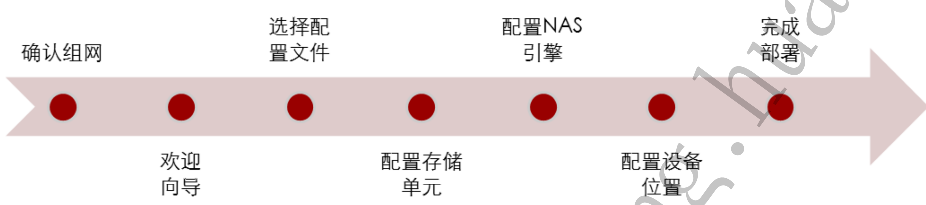
如图连接维护终端到管理网络，所有存储单元控制框（SUC）、集群NAS引擎节点（CNEN）的管理网口和维护终端连接到同一管理网络。维护终端配置一个与集群管理IP同网段的IP地址。



- 部署中QA问题
  - Q: 扫描时无法发现存储单元和引擎
  - A:开局工具使用了自定义的私有协议号发现设备，部分防火墙软件会对接收到的这些数据报进行屏蔽，请关闭防火墙后进行配置
  - Q:集群中显示的节点少于真实集群中的节点
  - A:可能原因集群中部分节点的管理网口未连接到网络或节点未正常上电
  - Q:同一套集群的两个节点分别显示在不同集群下
  - A:集群节点间的心跳未正确连接，导致集群内部无法发现其他节点
  - Q:配置时提示 IP 已经被使用
  - A:所配置的 IP 地址在网络中已存在，请更换为其他 IP 地址后重试
  - Q:命令设备初始化失败
  - A:在初始化的最后，会为系统创建带内命令设备及系统日志LUN，此时存储单元上为系统日志保留的磁盘若已经被使用，则该配置失败，请将保留的磁盘删除后重试
  - Q:设置 NAS 引擎基本信息并点击”下一步”后，提示 IP 无法连接
  - A:本机无法与所配置的物理 IP 连通，请检查网络配置后重试
  - Q:配置设备位置后并点击”下一步”后，提示位置信息未配置完全
  - A:位置信息(如机柜号，位置等)未配置完全，请正确配置所有设备的位置信息后重试

N8500部署流程

- 部署工具可以帮助工程师、运维工程师完成对NAS集群系统的部署任务
- 用户可以通过部署向导选择配置文件、配置存储单元、配置NAS引擎、配置机架位置等步骤，完成NAS一体化系统部署，部署完成后系统即可正常使用



Page 39
HUAWEI

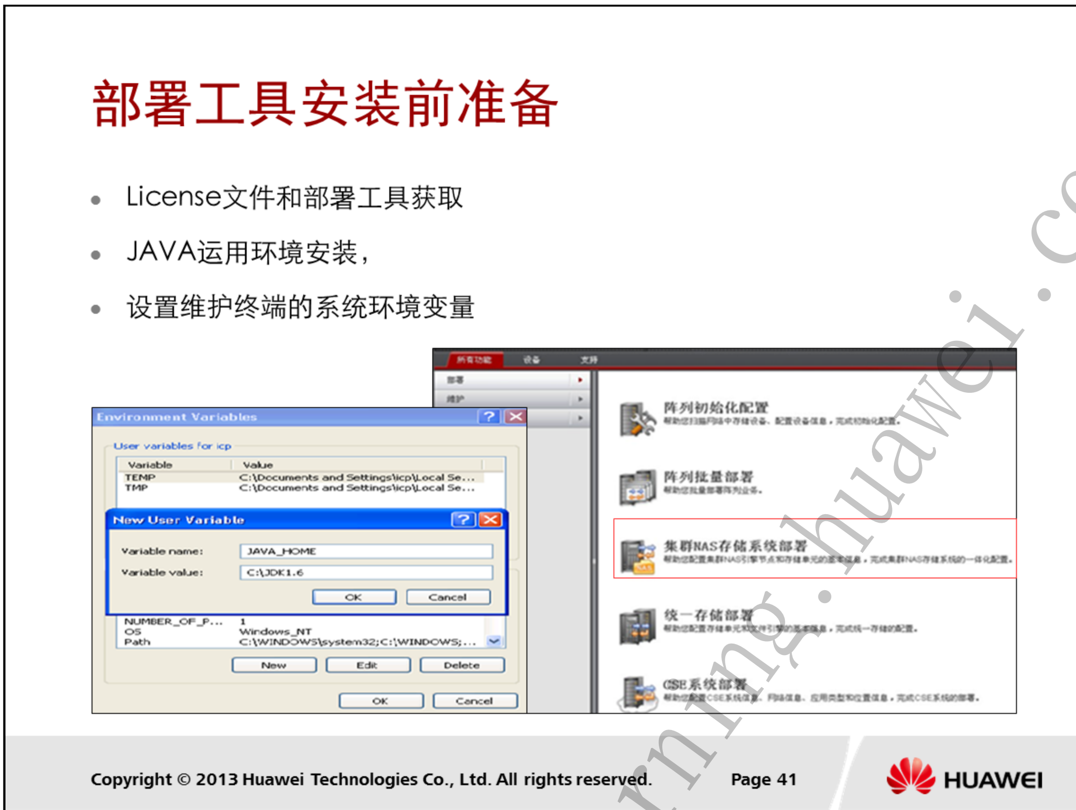
- Q:确认进行配置后初始化过程无法进行
- A:部署工具使用系统默认用户名和密码登陆，若所配置的存储单元或引擎已经被初始化且默认密码已经被修改，则登陆失败
- Q:部署工具无法正确运行
- A:请参考”部署工具安装前准备”，检查系统环境是否满足要求

部署工具环境要求

项目	详细信息
支持操作系统	Windows XP 32bit Windows 7 32bit Windows 2003 32bit Windows 2008 32bit Windows 7 64bit Windows 2003 64bit Windows 2008 64bit
JRE	JRE 1.6.0_30 32bit
内存	系统空闲内存要求400M以上。
硬盘空间	所在硬盘分区的剩余空间要求2G以上。
屏幕分辨率	1024*768及以上。
浏览器支持	IE7.0以上/FireFox4.0以上。

 Page 40 HUAWEI

部署工具软件包自带JRE，32bit和64bit操作系统上都统一以32bit 版本的JRE来支持运行Toolkit。



- 部署工具下载地址：http://support.huawei.com/enterprise/softdownload
- JAVA环境选择与JDK 1.6.0_30 或以上版本



- 在部署工具主界面中，单击“集群NAS存储系统部署”，进入组网图确认界面。
- 在部署工具主界面，点击“集群NAS存储系统部署”，进入“欢迎”向导界面，开始集群NAS的部署工作。



- 初始化配置有两种方式：
  - 导入配置文件，请确保模板文件的格式和配置信息正确。也可以点击“下载”链接，下载模板并填入正确的配置信息
  - 如果选择跳过此步骤，系统将提示手动输入配置信息



在“配置存储单元”界面，如果列表中没有相应设备，请确认设备与管理终端在同一子网且连接正常后，单击“扫描”按钮执行存储单元扫描功能，然后可以从同一子网中获取最新存储单元设备的信息 。

在选中您要初始化的存储单元后，如果您发现相应的存储单元信息不正确，您可以单击“修改”按钮对选中的存储单元进行信息修改。注意：每次只能修改一个存储单元。

在确认存储单元信息正确后，单击“下一步”；当出现警告对话框后选择“确定”；初始化操作后，若选中的存储单元的初始化状态有“失败”的，需要重新选择存储单元进行配置或不选择该“失败”的存储单元。



在“配置NAS引擎”界面，如果列表中没有相应设备，请确认设备在同一子网且连接正常后，单击“扫描”按钮执行NAS引擎扫描功能，然后可以从同一子网中获取最新NAS引擎设备的信息。

根据序列号，选中需要配置的集群，并选中其中之一，设置成集群的主节点，主节点NAS引擎support账户/密码默认为：support/support@storage（按照服务规范要求保证客户设备维护信息的信息安全）。



- 根据集群NAS引擎的节点IP地址规划配置每个节点IP地址：
  - 选择了导入配置文件，此处将会自动显示出配置文件中的IP地址的分配配置信息，此时您也可以对相关信息进行修改
  - 选择了手动输入配置信息，此时您可以手动对IP地址进行分配

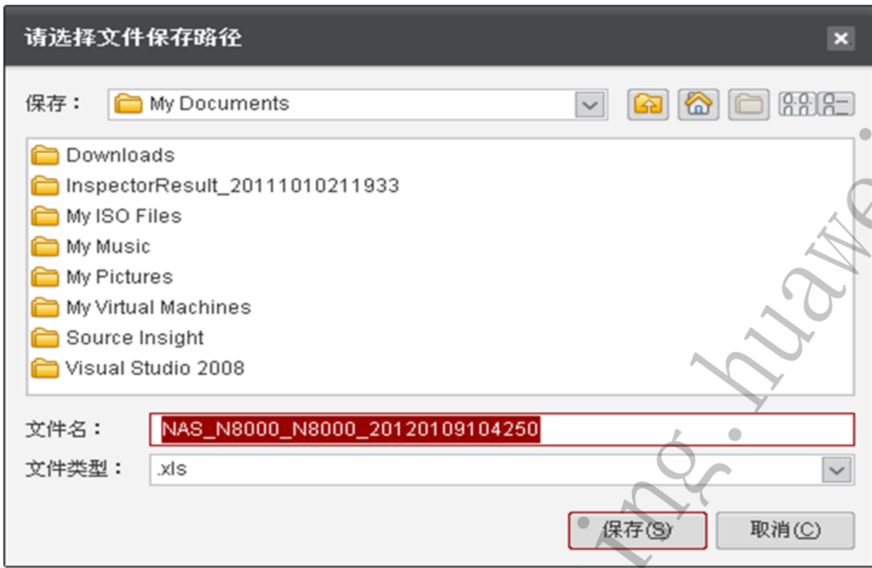


- N8500 license申请注意事项：
    - N8500通过License管理特性有：CIFS共享，NFS共享，FTP服务，动态分级存储，文件系统镜像，文件系统快，集群节点数
    - N8500集群节点数也受license控制，节点扩容后需要重新申请license，比如为2节点集群申请的license在系统扩容为4节点时不适用，需重新申请。
    - N8000从V1R2版本升级到V2R1版本，在不需要扩容的情况下，不需要重新申请License文件。
- 有关ESN
    - ESN（Equipment serial number）：设备序列号，是唯一标识设备的字符串，用来保证将License授权给指定设备的关键信息。
    - 集群NAS的ESN是由各NAS引擎每个框的ESN以(ESN1，ESN2，ESN3…)的形式组合而成，单个节点的序列号不能用于整个NSA产品的liecense申请。
    - 每框的SN位于框正面的右侧编码条上，由TYPE和S/N构成（单个ESN共20位，全部大写）
- 有关License授权许可证书
    - License授权许可证书随产品发货，可以用于查询合同号、客户PO、产品版本等信息。



- 有关License申请网站平台
  - ESDP平台的GTS权限（适用于原厂工程师）
  - 拥有GTS权限的用户，可以通过“授权激活”的方式申请License。如果尚未开通GTS权限，请单击ESDP平台首页左侧导航的“权限申请”，进入相应界面申请权限。
  - ESDP平台的ASP或Guest权限（适用于合作伙伴或最终用户）拥有ASP或Guest权限的用户，可以通过“密码激活”的方式申请License。请单击平台首页的“注册”，填写相关信息成为注册用户，即可自动获得权限。
- Licens申请具体操作，请参考N8500产品手册。

部署工具—配置文件保存



 Page 49 HUAWEI



在“启动部署”界面中，系统开始执行NAS引擎节点的添加，执行NAS引擎节点的集群加入，另外完成存储单元的带内初始化工作。完成以后便可以链接到ISM的下载页面，并且可以保存配置成功后的NAS集群信息。

Q: 扫描时无法发现存储单元和引擎

A: 开局工具使用了自定义的私有协议号发现设备，部分防火墙软件会对接收到的这些数据报进行屏蔽，请关闭防火墙后进行配置

Q: 集群中显示的节点少于真实集群中的节点

A: 集群中部分节点的管理网口未连接到网络或节点未正常上电

Q: 同一套集群的两个节点分别显示在不同集群下

A: 集群节点间的心跳未正确连接，导致集群内部无法发现其他节点

Q: 是否支持重新部署

A: 条件支持，需要登录已部署完成系统做一些清除操作。

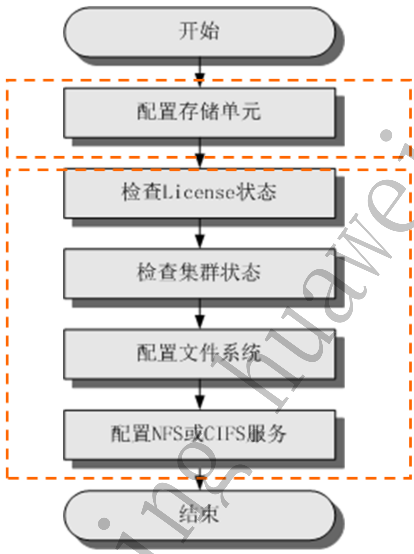
1. 登录到每一个节点(使用support帐户)上面，执行以下两个脚本
2. /opt/HS/scripts/clean_hs.sh
3. /opt/VRTSnasgw/scripts/cleanup.sh
4. 重启集群节点，然后重新执行开局工具进行部署。



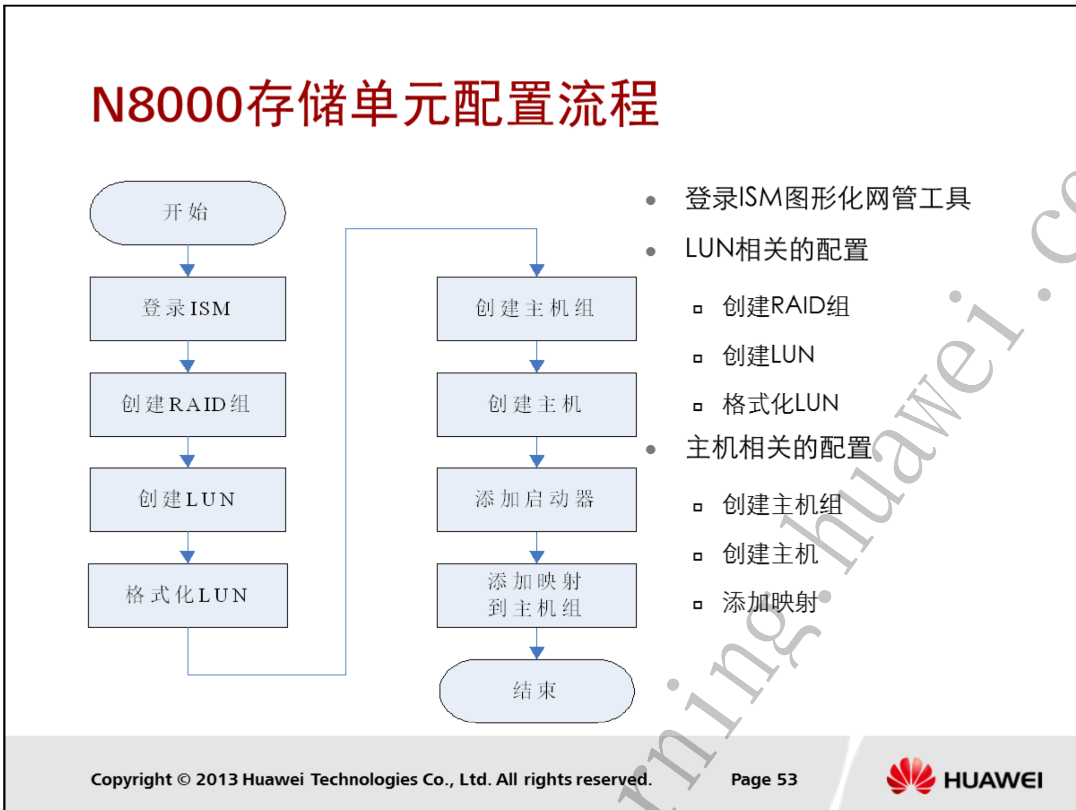
保存配置文件用于将集群的配置信息保存到维护终端上。在“启动部署”界面中，系统开始执行NAS引擎节点的添加，执行NAS引擎节点的集群加入，另外完成存储单元的带内初始化工作。完成以后便可以链接到ISM的下载页面，并且可以保存配置成功后的NAS集群信息。

N8000基本配置——总体流程

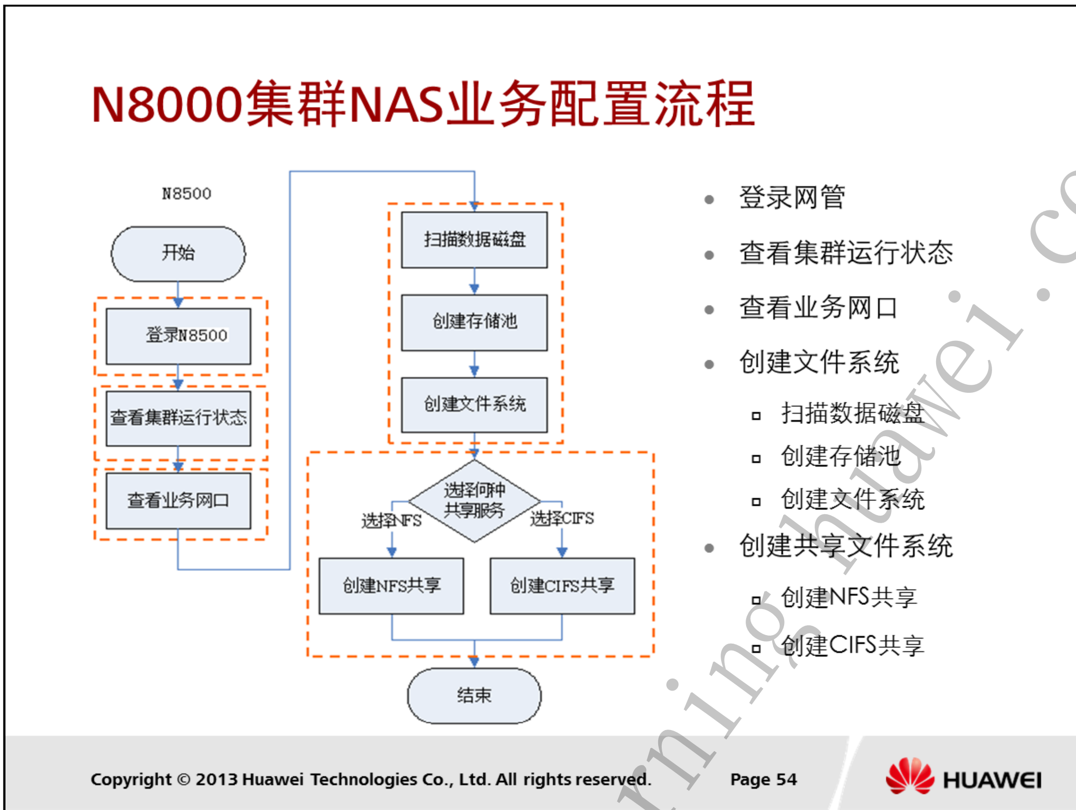
- 配置存储单元
- 配置NAS引擎基础业务
  - 检查License状态
  - 检查集群状态
  - 配置文件系统
  - 配置NFS或CIFS共享



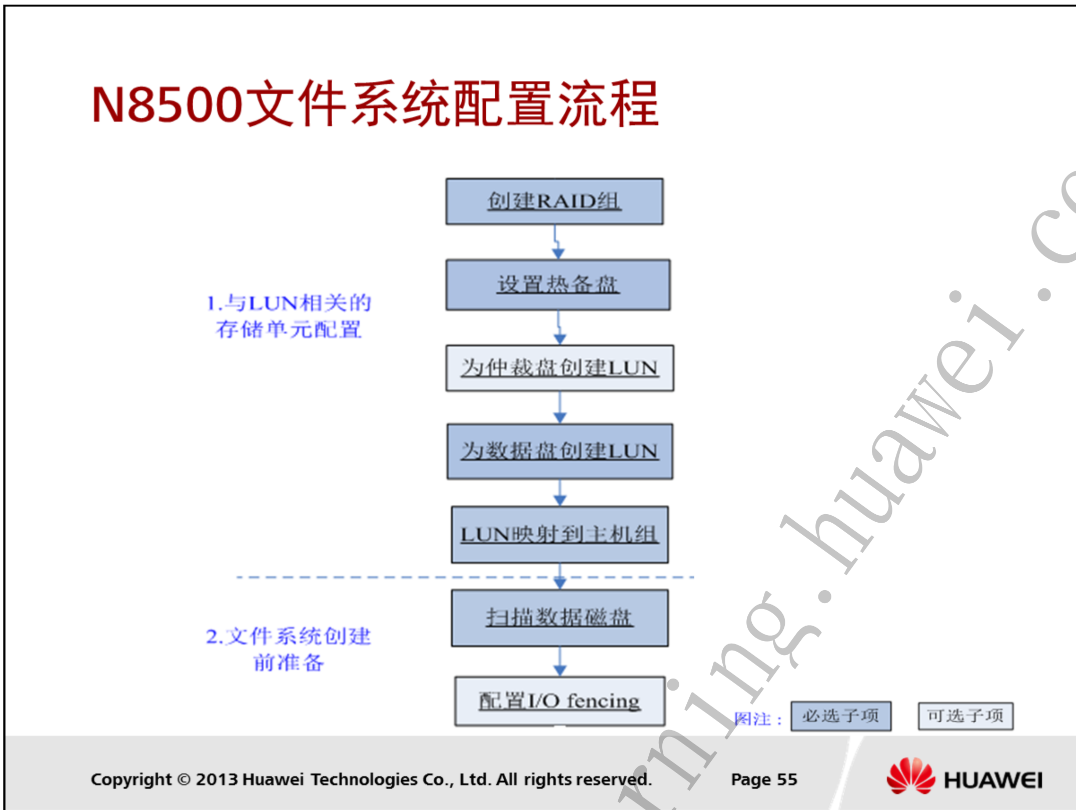
HUAWEI

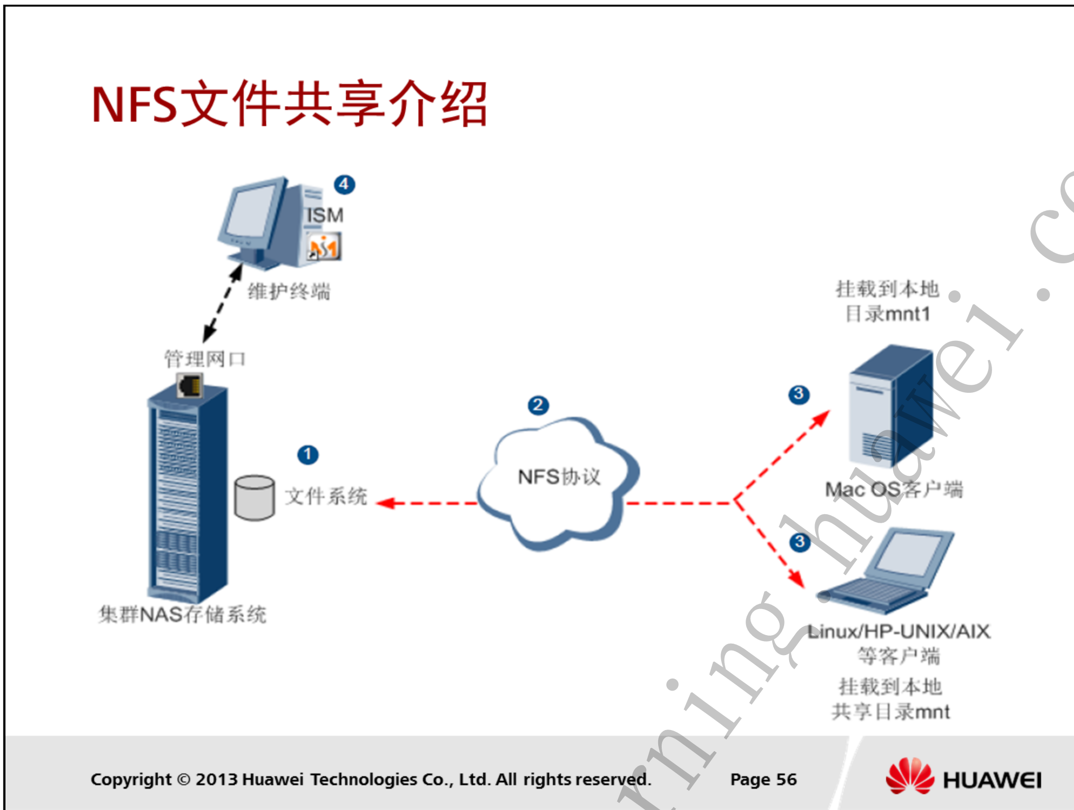


- 有关存储单元LUN的分条深度：
  - N8500应用于顺序IO为主的业务时，如存储媒体数据，建议设置较大的分条深度，如128
  - N8500应用于随机IO为主的业务时，如存储事务处理数据，建议设置较小的分条深度32或64
- 有关存储单元与NAS集群节点映射：
  - 存储单元侧创建一个主机组对应一个NAS集群，创建一台主机对应NAS集群中的一个节点
  - N8000初始化时已在存储单元侧为NAS集群创建了主机、主机组，因此可以重复创建

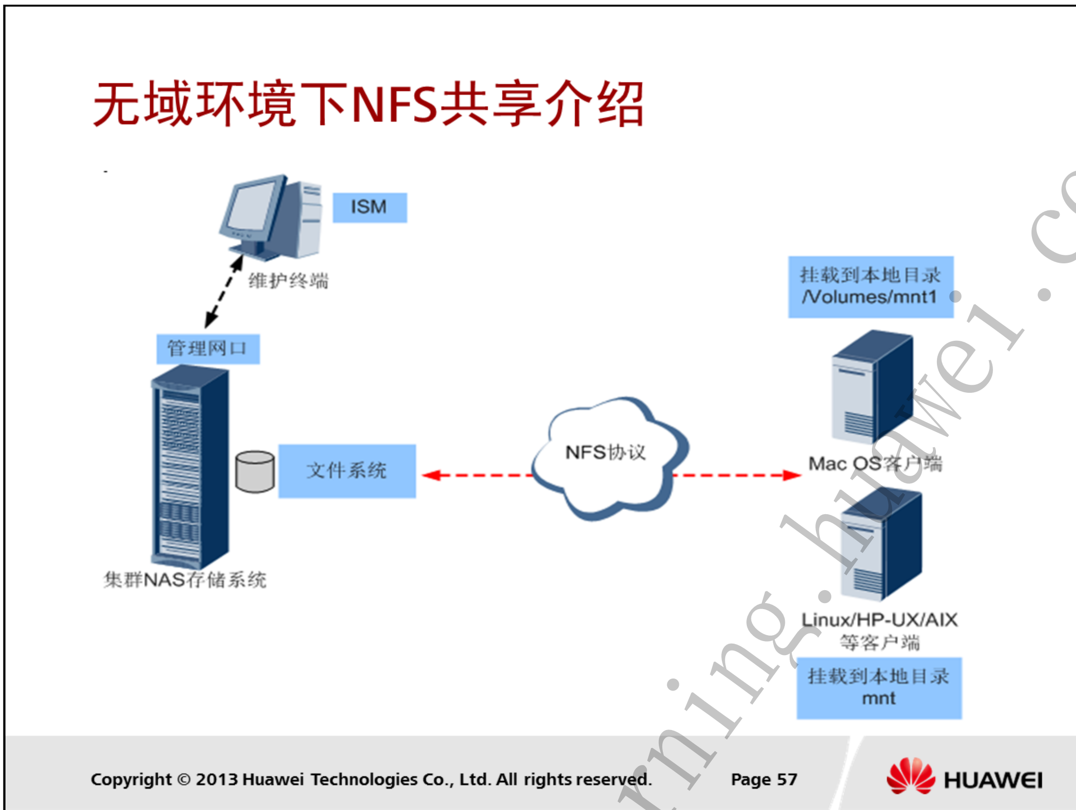


- 配置过程注意事项：
  - 集群中一个节点作为一个node，有一个WWN号，每个节点通过WWN来进行区分
  - 添加启动器选择标识（WWN号）时，通过WWN号的映射，将引擎节点和存储单元主机关联起来
  - 创建文件系统时，不要使用包含simple,stripped,mirror关键字为文件系统命名
  - 在创建镜像文件系统时，建议不要在同一阵列做镜像
  - 在创建文件系统时，不要使用"-"这个特殊字符
  - 创建文件时，文件系统名长度限制在3-12之间
  - 创建大于1T的文件系统时，选择块大小为8192B

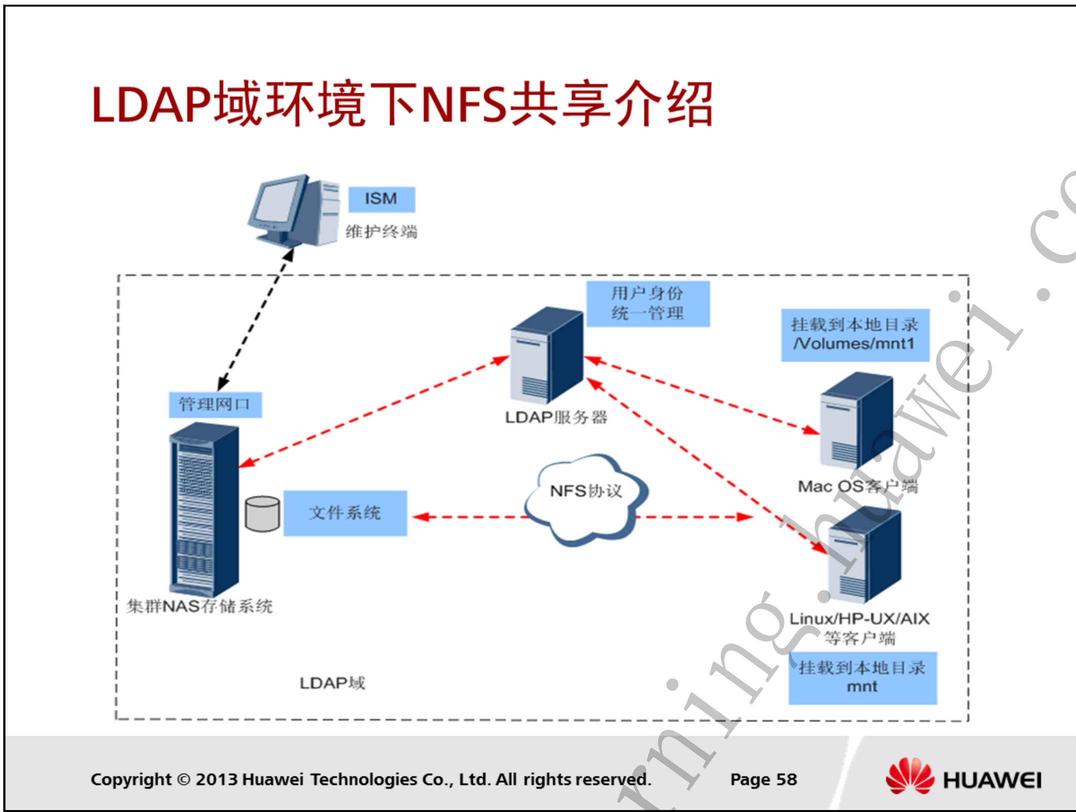

N8500文件系统配置流程
创建RAID组
设置热备盘
1.与LUN相关的
存储单元配置
为仲裁盘创建LUN
为数据盘创建LUN
LUN映射到主机组
扫描数据磁盘
2.文件系统创建
前准备
配置I/O fencing
附注：
必选子项
可选子项
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 55
HUAWEI



集群NAS存储系统主要由集群NAS引擎和后端存储单元组成，存储单元提供存储空间，集群NAS引擎提供文件系统。集群NAS存储系统将文件系统置于存储设备侧，在Linux/HP-UX/AIX/Mac OS等操作系统环境中，通过NFS协议将文件系统共享给客户端使用，系统直接处理文件访问请求，不再占用客户端资源，提供便捷、高性能价格比的文件存储。



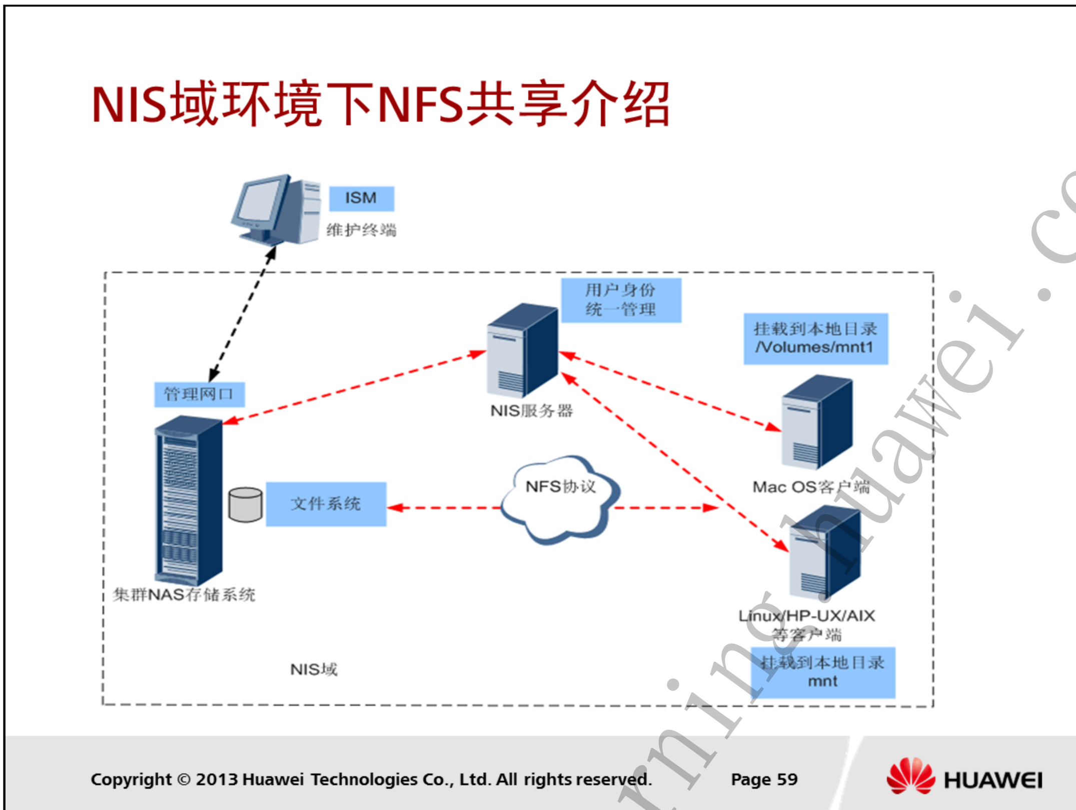
集群NAS存储系统作为服务器，通过NFS协议，向客户端提供对文件系统的共享访问，客户端挂载共享文件系统到本地，用户像访问本地文件一样远程访问服务器中的文件。在服务器侧通过设置客户端标识，确定允许访问该文件系统的客户端信息。



随着网络应用日益丰富，用户管理越来越复杂。对于提供单一服务的系统来说，基本上都是采用“用户名-密码”的认证方式。网络中的各种应用对每个用户具有不同的权限，这将导致每个应用都需要设定不同的用户名和密码。对于多个应用系统，用户需要输入不同的用户名和密码，不仅繁琐，而且不易管理。LDAP（Lightweight Directory Access Protocol）提供的目录服务能够有效解决此类问题。

LDAP是一个开放的、可扩展的网络协议。由于其简单、安全、优秀的信息查询功能，并且支持跨平台的数据访问，LDAP已逐渐成为网络管理的重要工具。基于LDAP的认证应用主要是实现一个以目录为核心的用户认证系统，即LDAP域环境。在LDAP域环境中，当客户端的用户需要访问应用程序时，客户端将用户名和密码提供给LDAP服务器，LDAP服务器将其与目录数据库中的认证信息进行比对来确定用户身份的合法性。

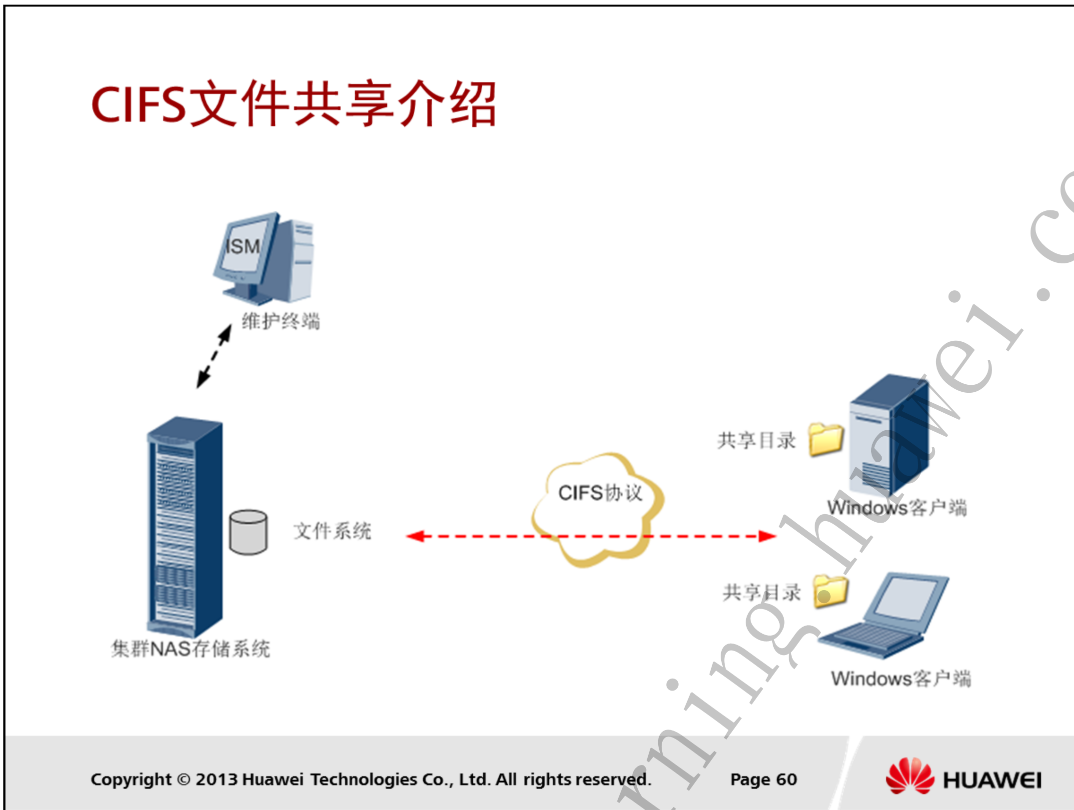
在集群NAS存储系统中，用户访问NFS共享时，将通过LDAP服务器完成用户身份的认证，保证文件系统数据可靠。



在一个独立应用的局域网系统中，如果不同的主机分别维护各自的网络信息，包括用户名、密码、主目录、组信息等，一旦网络信息需要更改，将是件非常复杂的事情。NIS（Network Information Service）是一种可以集中管理系统数据库的目录服务技术，提供了一个网络黄页的功能，能够有效解决此类问题。

NIS为网络中所有的主机提供网络信息。NIS使用客户端/服务器架构。如果某个用户的用户名以及密码保存在了NIS服务器中的数据库中，NIS允许此用户在NIS客户端上登录，并且通过维护NIS服务器中的数据库，可以统一管理整个局域网系统中的网络信息。

在集群NAS存储系统中，使用NIS进行访问控制。当用户访问NFS共享时，系统会到NIS服务器上去寻找用户使用的用户名和密码信息加以比对，以提供用户访问检验，保证文件系统数据可靠。

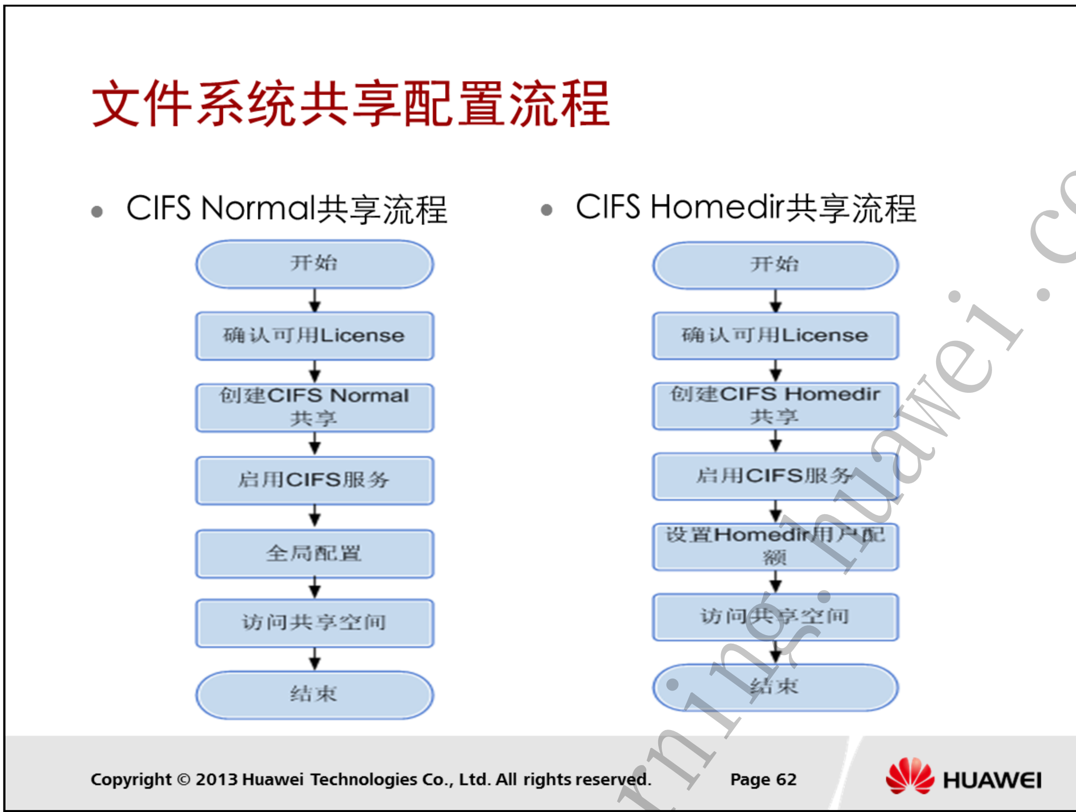


CIFS是一个网络文件共享协议，允许Internet和Intranet中的Windows主机访问网络中的文件或其他资源。N8000的CIFS特性使得Windows客户端能够识别并访问网络中集群NAS存储系统提供的共享资源，客户端用户能够像使用本机一样对保存在集群NAS存储系统中的文件进行读、写、创建等操作。

CIFS共享方式与全局模式

- N8000提供的CIFS文件系统共享主要应用于Windows客户端。
  - 共享方式分为CIFS Normal 和CIFS Homedir：
    - CIFS normal共享：如果不用区分限制访问共享目录的用户，可选择此共享
    - CIFS homedir共享：如果需要特定的用户拥有单独的用户目录，限制其他用户访问。
- 全局模式分为普通（Normal）模式及集群（ Cluster ）模式：
  - 普通模式：一个共享最终只会在一个节点上承载业务（但客户端可以通过任何一个VIP访问该共享）。
  - 集群模式：每个共享都会在所有节点上承载业务，该模式性能更好。

HUAWEI



Q：如何查看虚拟IP地址?

A：“全局模式”为“普通”：您可通过“所有设备 > 集群NAS存储系统 >NAS服务> CIFS共享 > 右侧操作区选中需要查看的共享名称 > 下方选择“虚拟IP地址”页签”查看用于访问该共享的虚拟IP地址。

“全局模式”为“集群”：选择“所有设备 > 集群NAS存储系统 > 设备信息”，选中集群NAS存储系统的设备名称。单击“网络管理”，选择“IP地址管理”，获取在线的虚拟IP地址。任意一个在线的虚拟IP地址都可以访问共享空间。

Q：重新启用CIFS服务时需要输入的密码是什么?

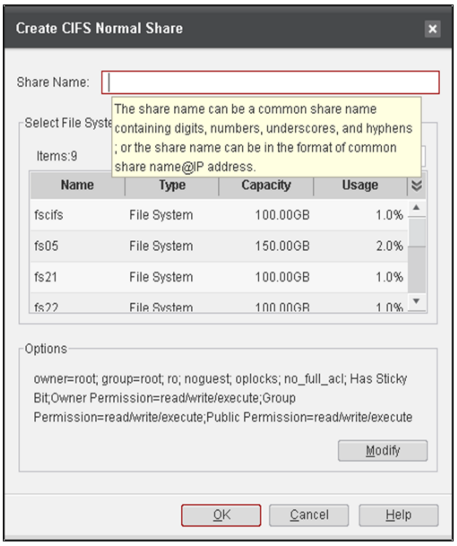
A：当集群NAS存储系统应用于AD域环境时，重新启用CIFS服务需要输入已加入AD域的域用户的密码。

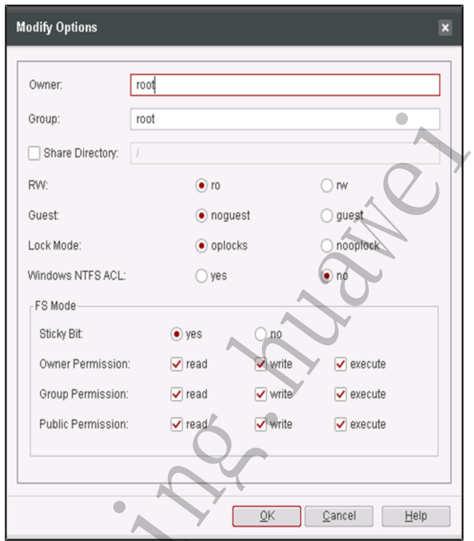
当集群NAS存储系统应用于无域环境时，重新启用CIFS服务不需要输入密码。

Q：通过不同的引擎IP地址是否可以同时读写同一共享文件?

A：可以同时读同一文件，当同时写同一文件时，由于集群有锁机制，只有一个IP地址能够写成功。

CIFS Normal共享中配置建议





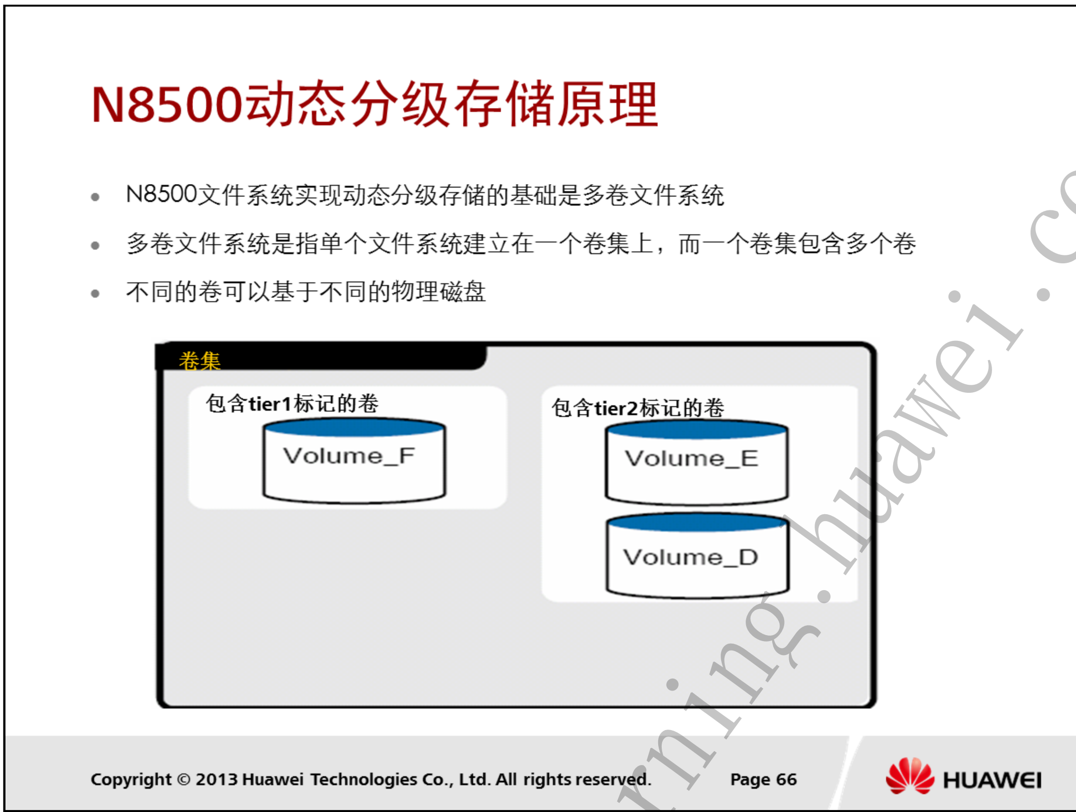
HUAWEI

目录

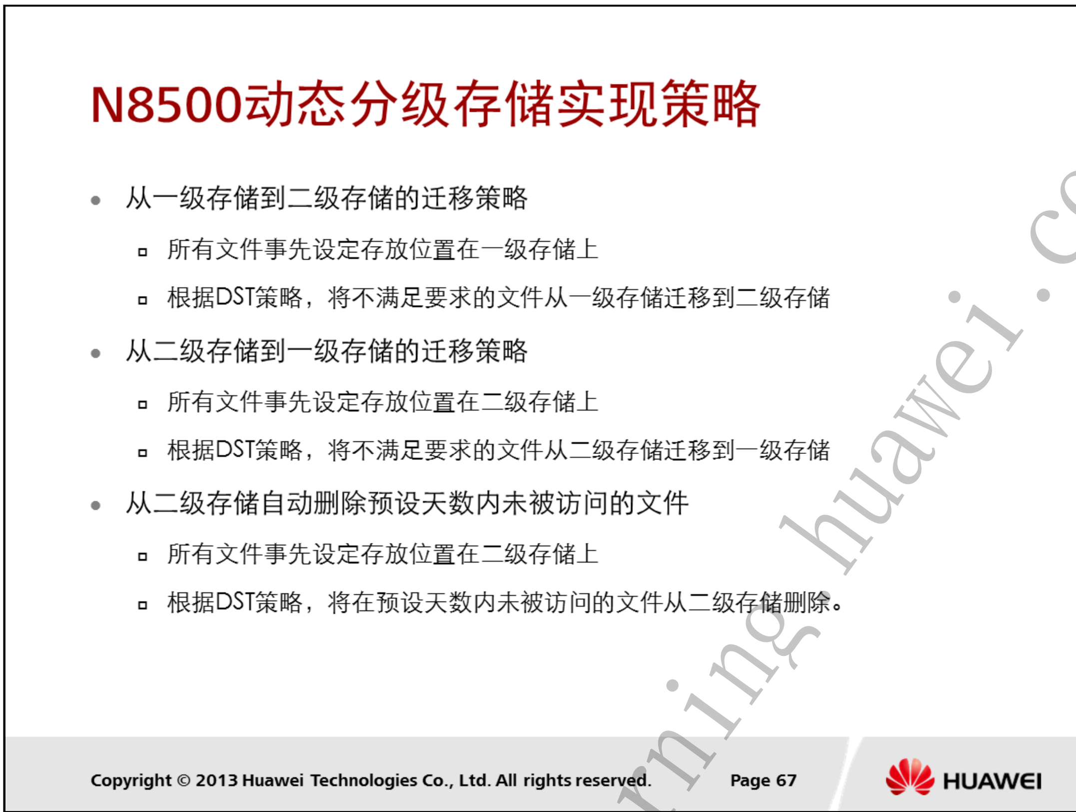
HUAWEI

N8500动态分级存储介绍

- 定义：
  - 动态分级存储（DST, Dynamic Storage Tiering），是指根据制定的存储策略，将访问频率较低的的数据迁移至成本较低的存储资源上，将经常访问的重要信息（即热点数据）存放在高性能存储设备上。
  - 当存放在廉价存储设备上的数据访问热度升高时，数据可自动迁移至高性能设备上。该迁移过程是自动、透明的。



图中表示了一个N8500多卷文件系统，该系统的卷集由分别名为 tier1、tier2的放置级别组成。这样的卷集可用于包含少量重要文件 (tier1)、大量非活动的文件 (tier2) 的文件系统。 多卷文件系统对于应用程序是透明的。该文件系统中的所有文件都是同一命名空间的一部分。通过多卷文件系统，可以实现同一文件系统的文件，可以存储在不同级别的存储介质上，如：热点数据和普通数据分别存放在SSD盘和SATA盘上。

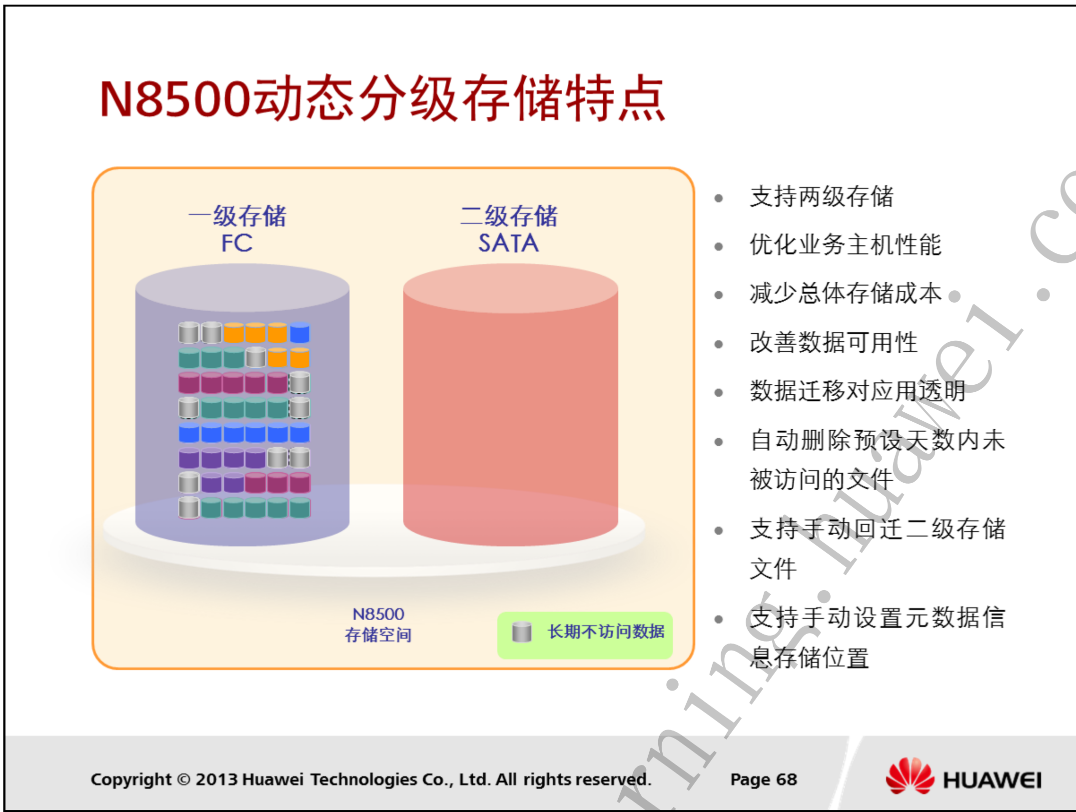


文件在指定时间(以天为单位)内没有被访问或者修改，文件就满足从一级到二级的迁移条件。迁移动作通过定时或者手工执行策略执行。

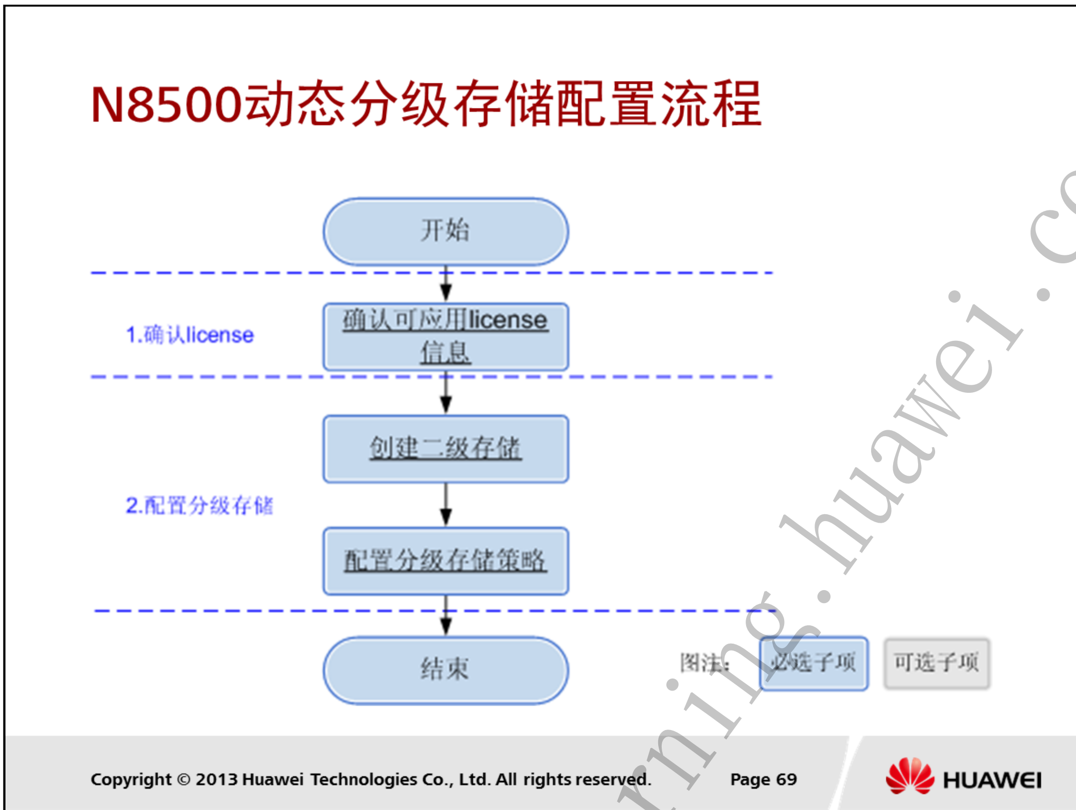
文件在指定周期内的平均IO访问次数超过设定的最小阈值，文件就满足从二级到一级的迁移条件。迁移动作通过定时或者手工执行策略执行。文件在指定周期内的平均IO访问次数被称为文件访问热度。操作系统内核中一个系统调用(读写)被记为一次IO访问。

二级存储通常会选择性价比较高的磁盘作为存储介质，如SATA磁盘。相对的，一级存储通常创建在性能较高，价钱也相对昂贵的磁盘上，如FC磁盘。

一级存储和二级存储的空间大小可以设置为不一样。



- 支持两级存储：N8500支持两级不同类型的存储。
- 优化业务主机性能：热点数据从高性能存储访问，提高系统性能；可设置热点访问策略，根据二级存储上数据的访问热度，把数据回迁至一级存储，实现性能最优化。
- 减少总体存储成本：根据策略筛选文件，将长期不访问的数据迁移到廉价磁盘上，可综合发挥不同存储设备的成本优势。
- 改善数据可用性：动态分级存储把很少使用的历史数据迁移至成本较低的存储资源中，避免了反复保存，减少了存储的时间；同时提高了在线数据的可用性，使磁盘的可用空间维持在系统要求的水平上。
- 数据迁移对应用透明：基于信息生命周期管理策略（ILM），无须人工干预。将数据迁移至另外的存储器时，应用程序不需要改变，使数据迁移对应用透明。



- 配置过程常见问题FAQ：

  Q：文件在两级存储之间反复迁移?

  A：检查迁移策略：文件同时满足两个迁移条件。

  Q：执行手动回迁之后的文件仍然在二级存储上?

  A：检查文件系统对应的迁移时间表：回迁动作执行时正好是迁移时间表的执行时间，回迁之后的文件又被迁移到了二级存储。

  Q：迁移执行之后，文件只有一部分迁移到了二级存储?

  A：检查文件系统二级存储空间：二级存储空间已满。

  二级存储上碎片过多。

  Q：迁移过程中，fsppadm进程占用CPU过高?

  A：fsppadm进程占用CPU情况与文件系统的块大小、文件系统中的碎片情况以及此次需要迁移的文件个数相关。建议在执行迁移之前对文件系统进行碎片整理，并在业务量不大时执行迁移策略。

创建二级存储ISM配置



 Page 70 HUAWEI

1. 在ISM主界面的导航树上选择“集群NAS存储系统 > 存储资源 > 文件系统”
2. 选择需要创建二级存储的文件系统，并在管理窗口下方分别单击“一级存储”和“二级存储”页签，查看当前文件系统的信息。
3. 在右侧操作区中，单击“分级存储策略管理 > 创建二级存储”。 系统弹出“创建二级存储”对话框。
4. 单击“下一步”。 系统弹出“创建二级存储 步骤3-1：设定容量”对话框。
5. 设定容量
6. 在列表框中，选择存储池（如果存储类型为数据磁盘，则选择数据磁盘）。
7. 单击“下一步”。 系统弹出“创建二级存储 步骤3-2：设定布局类型”对话框 。
8. 设定布局类型
9. 单击“下一步”。 系统弹出“创建二级存储 步骤3-3：信息汇总”对话框。
10. 单击“完成”。 系统弹出“执行结果”提示框，等待2分钟左右。
11. 单击“关闭”完成操作。
12. 在文件系统列表中选择添加了二级存储的文件系统，并在管理窗口下方分别 单击“一级存储”和“二级存储”页签，查看该文件系统二级存储是否添加成功



1. 在ISM主界面的导航树上选择“集群NAS存储系统 > 存储资源 > 文件系统”。 在右侧操作区选择需要配置分级存储策略的文件系统。
2. 在右侧操作区单击“分级存储策略 > 分级存储策略管理”。 系统弹出“分级存储策略管理”对话框。
3. 在“分级存储策略”区域框中，单击“修改”。 系统弹出“修改策略”对话框，修改当前文件系统的策略。
4. 设置策略执行时间表。在“分级存储管理”对话框，单击“时间表”区域框中的“创建”。 系统弹出“创建时间表”对话框。
5. 执行分级存储策略。 有两种方法可以实现，第一种是根据执行时间表自动执行；第二种是手动执行策略。如果选择手动执行策略，在“分级存储管理”对话框中，单击“执行策略”。运行完成后，查看一级和二级存储中的文件是否发生了迁移。

目录

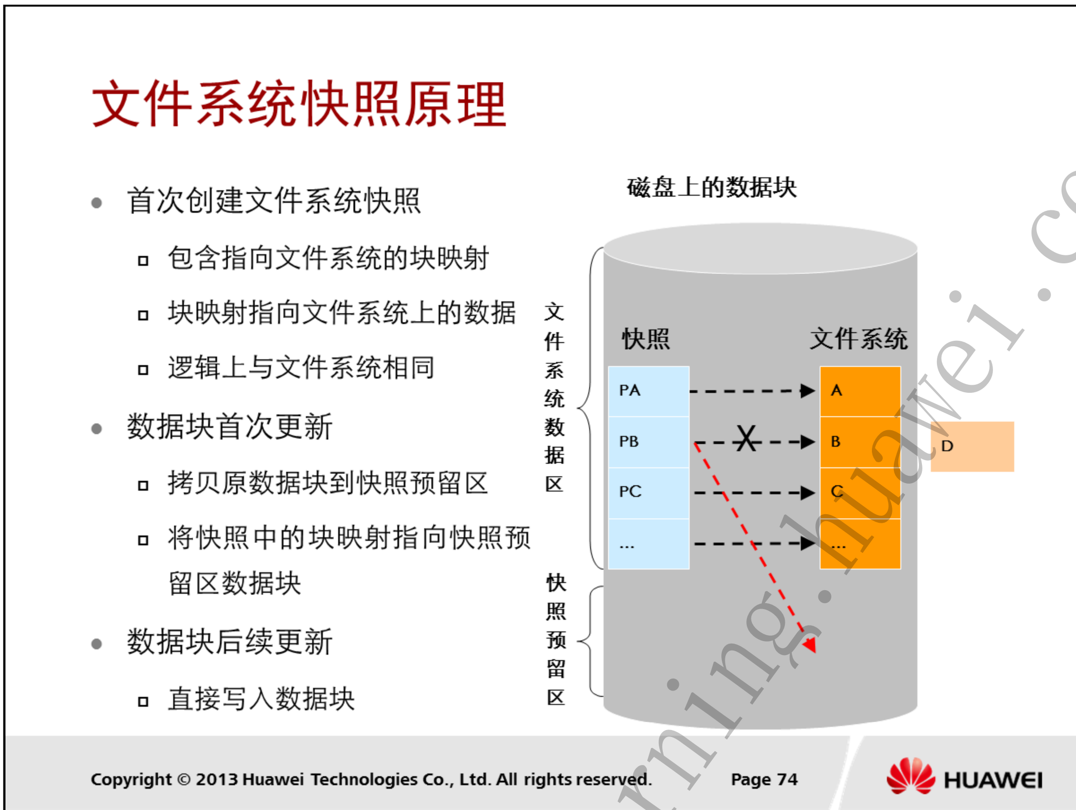
N8500文件系统快照的定义

- 定义：
  - 文件系统快照是文件系统的即时点精确映像
  - 文件系统快照为文件系统提供备份机制
  - 文件系统快照是与文件系统完全一致的视图
  - 存储检查点（Checkpoint）是文件系统管理和控制的对象
  - N8500文件系统快照是基于存储检查点（Checkpoint）的快照

Page 73
HUAWEI

- N8500文件系统快照是NAS服务中的增值功能

1. 文件系统快照是在某一时刻生成的文件系统即时映像，可以对文件系统提供备份机制。
2. 快照与生成快照时刻的文件系统是完全一致的视图。在生成快照之后，即使文件系统发生了改变，快照仍然保持原有的数据。
3. 存储检查点相当于是一系列指向文件数据块的指针，可以快速创建，占用少量的存储空间。
4. N8500的文件系统快照是基于存储检查点的快照，因此，具有存储检查点的特性。



- N8500文件系统快照原理解分为三个场景理解:

1、首次创建文件系统快照:

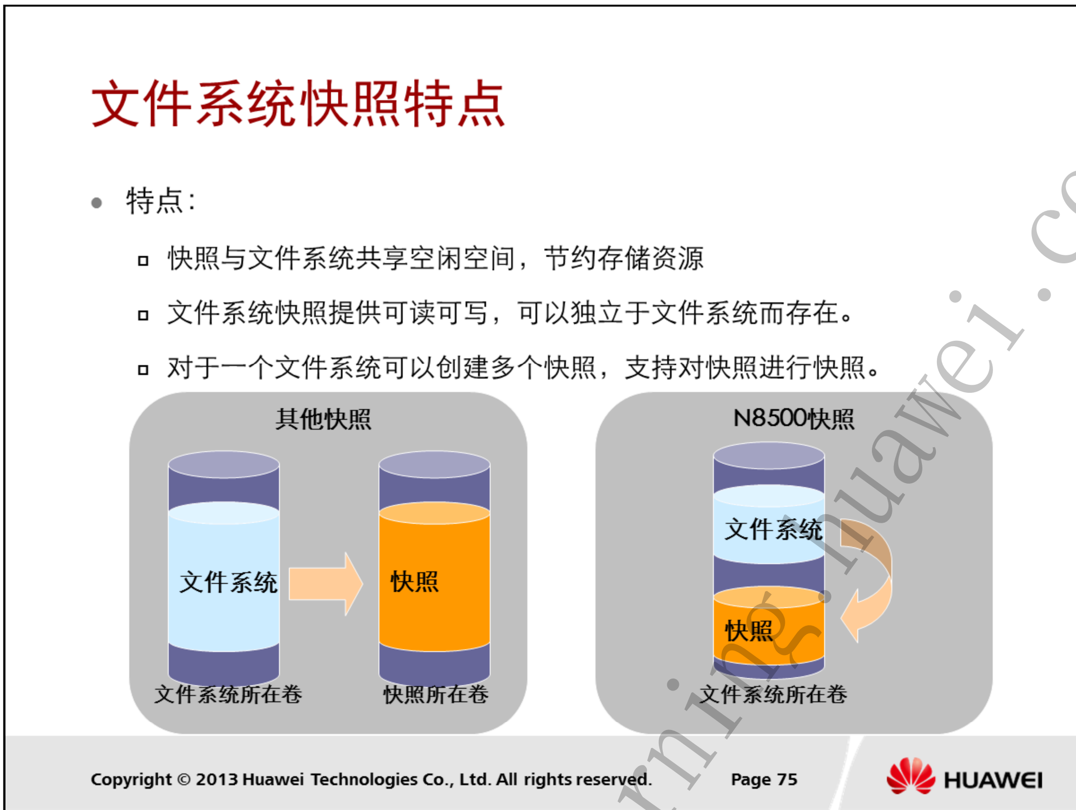
- 创建存储检查点，生成指向文件系统数据块的块映射。
- 完成快照的创建后，读取快照时将通过块映射直接获取文件系统里面的数据块。因此，快照与文件系统逻辑上是相同的。

2、文件系统数据块首次更新:

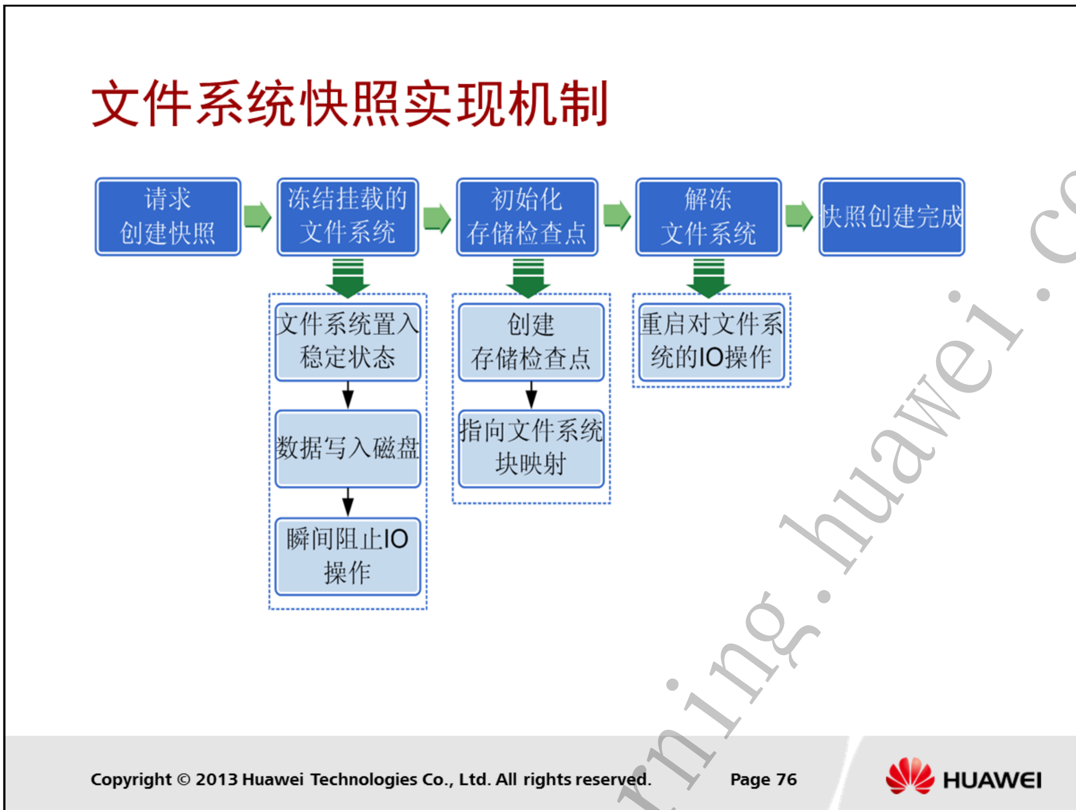
- 文件系统数据块需要更新时，为了保证读取快照的时候，仍然访问到之前的数据，需要进行以下步骤。
- 拷贝需要更改的数据块的原数据到快照预留区
- 取消原有块映射，并生成新的指向快照预留区的块映射。
- 更改数据块

3、文件系统数据块后续更新:

- 如果需要再次修改数据块，如图例中的数据块B`，只需要直接将新数据写入即可。这样的机制被称为Copy on Write（写时复制）。



- N8500使用的基于存储检查点的快照特点如下：
  - N8500快照与文件系统共享空闲空间，无需像其他快照一样需要为快照分配独立的存储空间，最小化对磁盘空间的使用。
  - N8500快照不仅向其他快照一样提供只读访问，还可提供写访问。在N8500应用中，能够被当作单独的空间共享出去，为客户端使用。此外，与其他快照不同的是，N8500快照不是完全依赖于文件系统，当文件系统unmount后，仍然可以被访问。
  - 对同一个文件系统可以创建多个快照。通过识别和维护自上次更改存储检查点或备份以来，使用写入时复制技术更改的文件系统块，存储检查点可极大减少 I/O 开销。有多个只读存储检查点，可减少 I/O 操作和所需存储空间，因为最新存储检查点是仅来自主文件系统的累计更新。



- 为了实现N8500快照的实现过程，主要分三个步骤：

1、冻结挂载的存储系统

- 目的：临时阻止对文件系统的IO操作，保证创建的存储检查点创建完成后保持与文件系统的一致性。
- 操作步骤：1）将文件系统置入稳定状态；2）将内存中的数据写入磁盘；3）瞬间阻止IO操作

2、初始化存储检查点

- 目的：完成存储检查点的创建，并保证该存储检查点是指向文件系统的。
- 操作步骤：1）创建存储检查点；2）指向文件系统块映射

3、解冻文件系统

- 目的：是文件系统在完成快照创建后，允许IO操作。
- 操作步骤：重启对文件系统的IO操作
- 注意事项：基于存储检查点的快照时间比较短，一般几分钟，具体时间取决于文件系统中文件的数量多少。

文件系统快照分类

手动快照

- 需要手动指定文件系统
- 可以在任意时刻创建快照
- 可以为手动快照命名
- 创建完成的快照状态默认显示为“Online”

自动快照

- 需要根据定时策略创建快照
- 可手动指定自动快照的数目，每个文件系统最多不超过128个
- 自动快照命名规则为“策略名称”+“日期”+“月份”+“年”+“时刻”+“时区”。
- 创建完成后默认状态显示“Offline”

手动快照与自动快照

- 每个文件系统的手动快照与自动快照加起来的总数不能超过512个
- 状态处于“Online”的快照文件系统一样被共享出去

 Page 77 HUAWEI

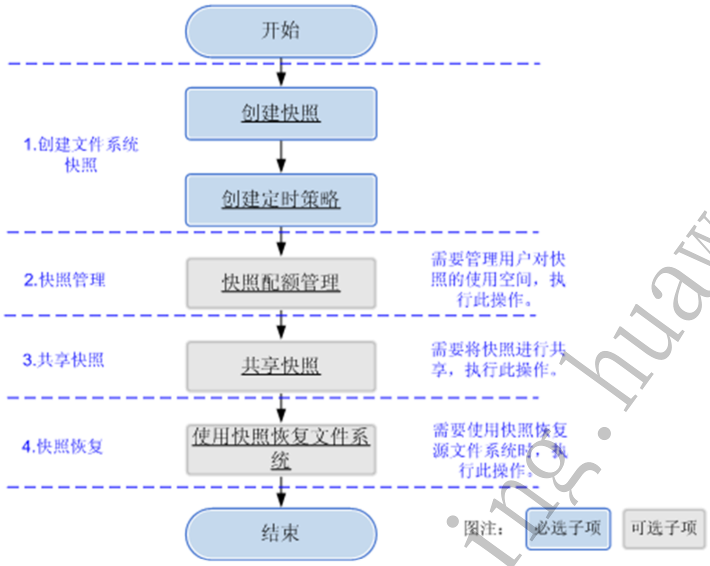
快照配置

- 配置前提：
  1. N8500系统中已创建一个或多个文件系统。
  2. 需要创建快照的文件系统剩余空间充足。
  3. 当前文件系统已有快照，但快照容量未达到快照配额限制。
  4. 需要创建快照的文件系统快照数目未达到限额。

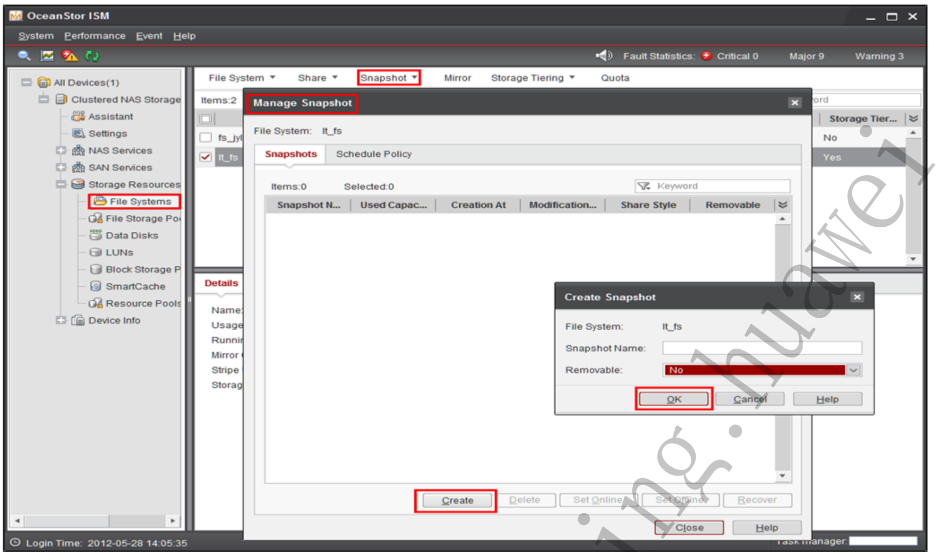
 Page 78 HUAWEI

- 文件系统快照配置前提：
  - 快照是基于文件系统的，在创建快照时必须指定创建哪个文件系统的快照。因此，文件系统的创建是快照创建的前提。
  - N8500快照与文件系统共用空间，因此需要创建快照的文件系统必须有保留有存储快照的空间。
  - N8500提供快照配额管理，可手动设置快照所占用的存储空间。如果超过快照配额，创建快照将不会成功。
  - N8500单个文件系统的快照最大数量为512个，单个文件系统的定时快照最大数量为128个。

快照配置流程


开始
创建快照
1.创建文件系统快照
创建定时策略
2.快照管理
快照配额管理
需要管理用户对快照的使用空间，执行此操作。
3.共享快照
共享快照
需要将快照进行共享，执行此操作。
4.快照恢复
使用快照恢复文件系统
需要使用快照恢复源文件系统时，执行此操作。
结束
图注：
必选子项
可选子项

N8500产品手动快照ISM配置



 Page 80 HUAWEI

- 文件系统-> 快照-> 快照管理-> 创建

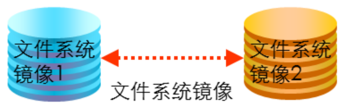


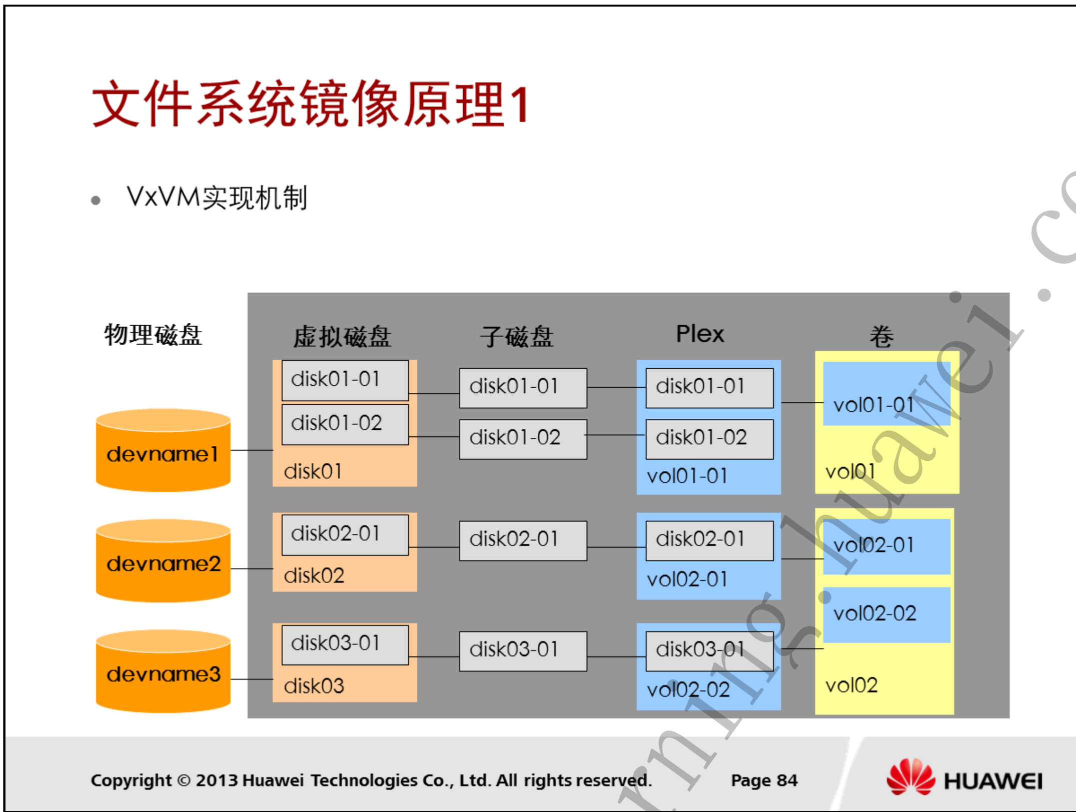
- 文件系统-〉快照-〉快照管理-〉定时策略-〉创建
- 创建快照时文件系统的IO会被冻结，故所有业务会中断。建议在业务量小或者无业务时进行快照创建。
- 快照依赖于文件系统而存在，文件系统删除的同时，快照也被删除。

目录

N8500文件系统镜像介绍

- 定义：
  - 文件系统镜像，是在两个或多个磁盘或磁盘子系统上产生同一个文件系统镜像视图的信息存储过程。
  - N8500 支持多重镜像，多重镜像是指一个文件系统拥有不止一个镜像，从而增强了数据的可靠性。
  - N8500对于单个文件系统最多支持3个镜像。



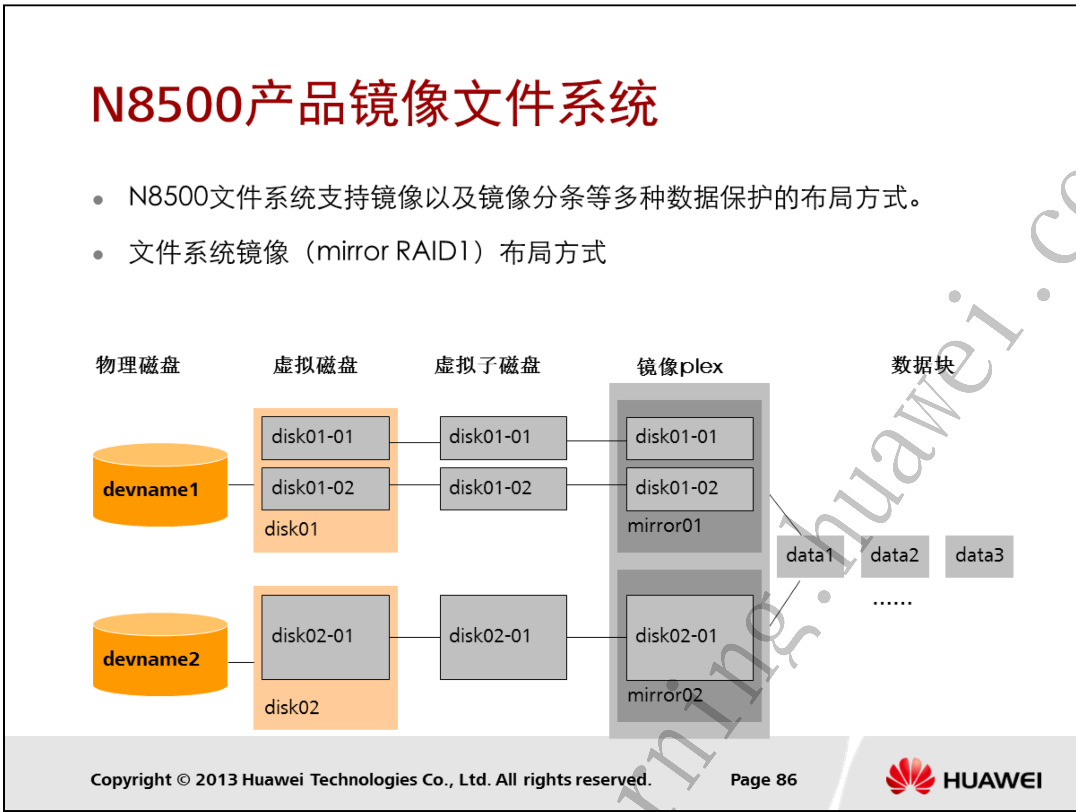


- 文件系统镜像的实现方式是基于文件系统使用的存储资源是经过多层虚拟化的。
  - 物理磁盘：物理磁盘是基本的存储设备（介质），即最终存储数据的地方。如：devname1
  - 虚拟磁盘：虚拟磁盘是基于同一个物理磁盘虚拟出来的逻辑单元。如：disk01
  - 子磁盘：子磁盘是一组连续的磁盘块。块是磁盘上的空间单元。N8500集群NAS软件存储资源管理使用子磁盘分配磁盘空间。一个虚拟磁盘可以分成一个或多个虚拟子磁盘。如：disk01-01，disk01-02
  - Plex： N8500集群NAS软件存储资源管理使用子磁盘构建称为 plex 的虚拟对象。Plex 由位于一个或多个物理磁盘上的一个或多个子磁盘组成。
  - 卷：卷是一种虚拟的磁盘设备，就应用程序、数据库和文件系统而言，它与物理磁盘设备很相似，但没有物理磁盘设备的物理局限性。卷由一个或多个 plex 组成，每个plex 都包含卷中所选数据的一个副本。



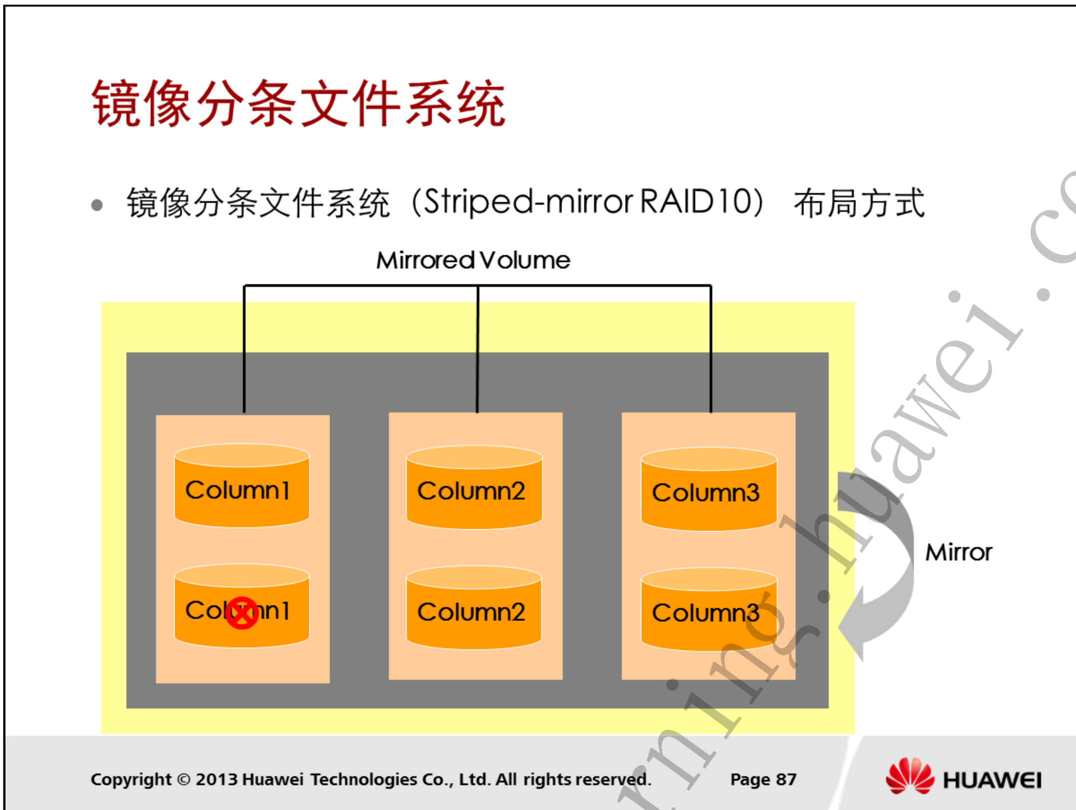
镜像分条（Striped-Mirror）RAID 10，在镜像的基础上进行条带化，兼有条带化和镜像的双重优点，即在利用镜像提供数据冗余的同时可以跨多个磁盘分布数据。

分条镜像（Mirrored-Stripe ）RAID 01，在条带化的基础上进行镜像，兼有条带化和镜像的双重优点，即在用镜像提供数据冗余的同时可以跨多个磁盘分布数据。



如果文件系统选择镜像布局，则可以看到一个Plex中包括至少2个镜像，每个镜像包含分属不同虚拟磁盘的子磁盘。当数据块写入时，同时写入两个镜像，从而最终同时写入两个不同的物理磁盘。

文件系统在写入数据时，N8500将数据同时写入到多套阵列的磁盘中，而且写入内容完全一致，这等于彼此备份。当其中一套阵列的数据出现故障时，可以马上从另一个阵列中进行恢复。注意，这两份数据并不是主从关系，而是相互镜像/恢复的。



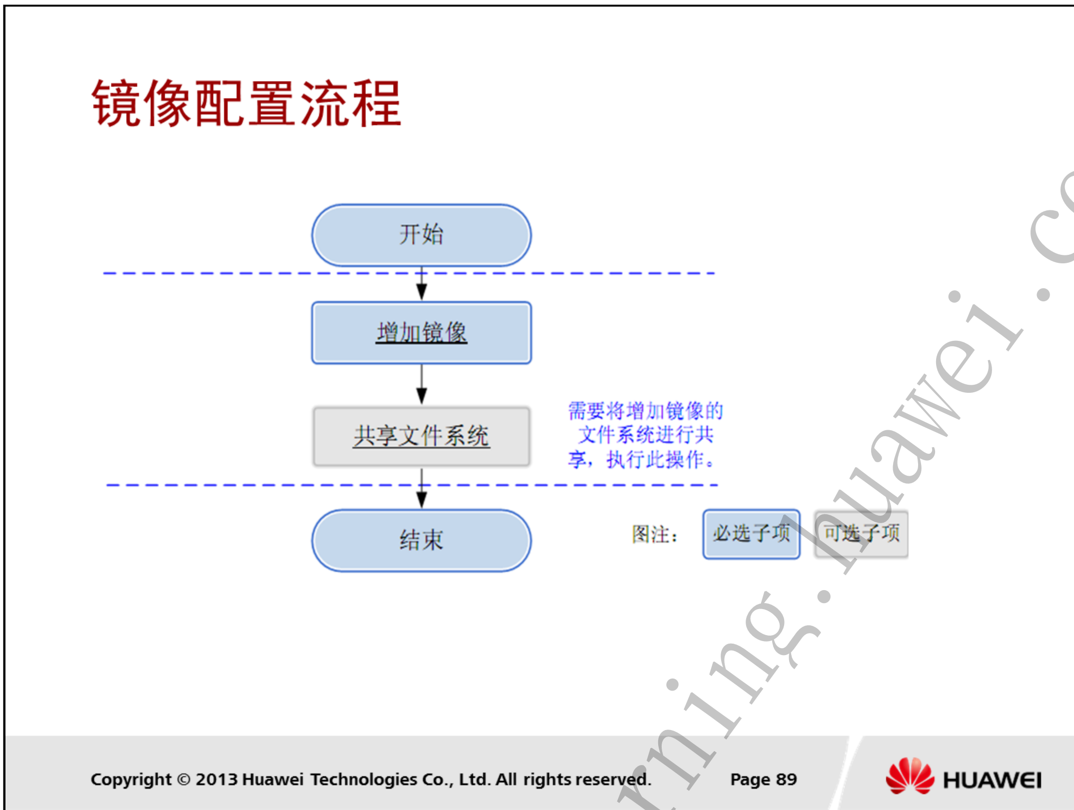
N8500文件系统支持在条带的基础上实现镜像。镜像分条布局兼有条带化和镜像的双重优点，即在利用镜像提供数据冗余的同时可以跨多个磁盘分布数据。它可以增强冗余并缩短磁盘发生故障后的恢复时间。

从图中可以看出，镜像分条布局在底层先进行了镜像布局，然后再外层进行了分条布局。当底层镜像卷中的某个column的镜像失效时，如Column1的一个镜像失效，那么只有Column1 的镜像被破坏，其他Column 仍然有独立的镜像。从整个Striped-mirror Volume仍然是部分冗余的。数据可靠性得到了保证。

N8500文件系统镜像配置准备

- 阵列间镜像配置前提：
  - 准备至少两套存储单元。
  - 存储单元中的磁盘类型建议为相同类型，如SATA，SAS或FC盘。
  - 两套存储单元建议配置成相同的布局，比如配置相同数目的RAID组，每个RAID组下有相同数目的LUN，每个LUN的大小与另一套阵列一致。
  - 建议将两套存储单元映射到引擎的LUN划分到不同的存储池中，便于管理和文件系统创建时选择不同来源的存储资源。

HUAWEI



- N8500文件系统镜像的配置过程包括：
  - 配置存储单元1：为创建文件系统提供存储资源。
  - 配置存储单元2：与配置第一套存储单元的步骤基本相同，只是在LUN相关配置中，无需为IO fencing功能创建对应的仲裁盘。因为仲裁盘只需要在第一套存储单元中创建了即可。
  - 创建文件系统镜像时可以选择创建镜像文件系统也可以选择创建镜像分条文件系统。其中创建存储池和创建文件系统是关键步骤。
    - 在创建存储池时，尽量将不同的阵列的LUN划分到不同的存储池；
    - 创建镜像文件系统，需要选择与镜像个数相同或者大于镜像个数的数据磁盘。选择的数据磁盘需要来源于不同的存储单元。
    - 创建镜像分条文件系统：同样需要选择不同阵列中的磁盘作为文件系统的镜像。



- 文件系统镜像配置FAQ：

Q：添加镜像时选择与当前文件系统大小相同的pool，为什么提示空间不足？

A：为了实现快速重同步，在添加镜像时会创建DCO卷，DCO卷需要额外的空间。

Q：如何完成镜像同步？

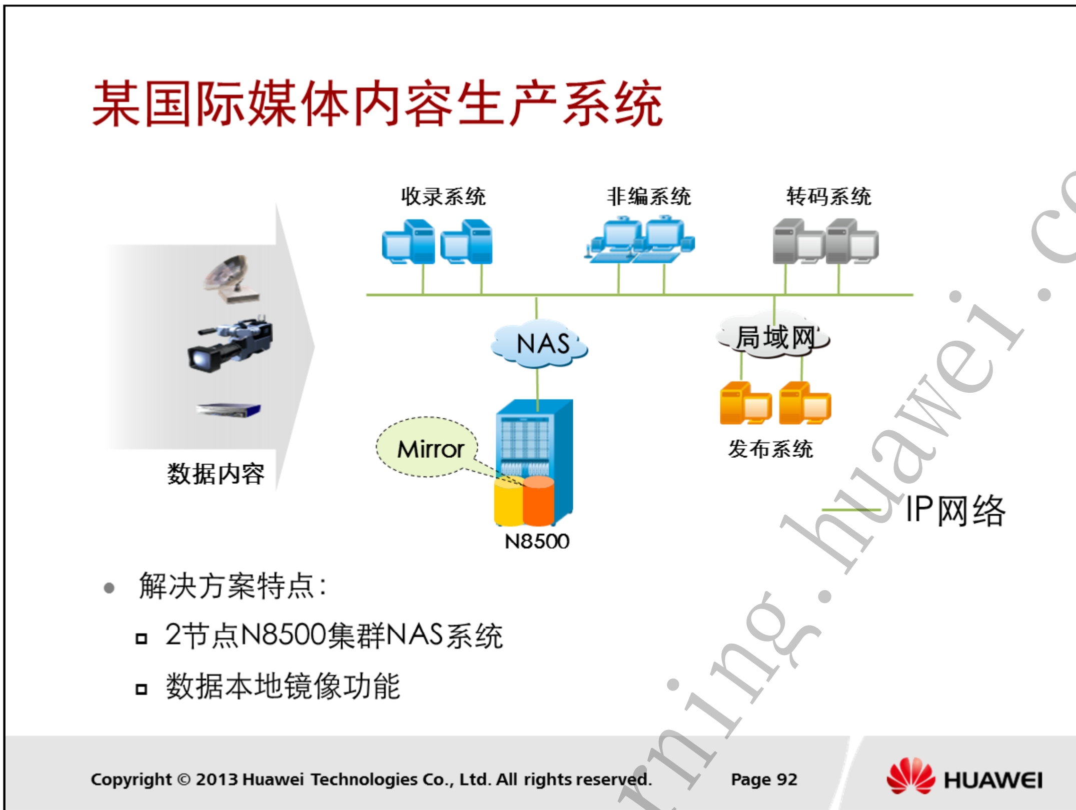
A：master用户ssh登录集群控制台IP，执行命令storage fs resync。

Q：N8000卷镜像功能是同步方式还是异步方式？

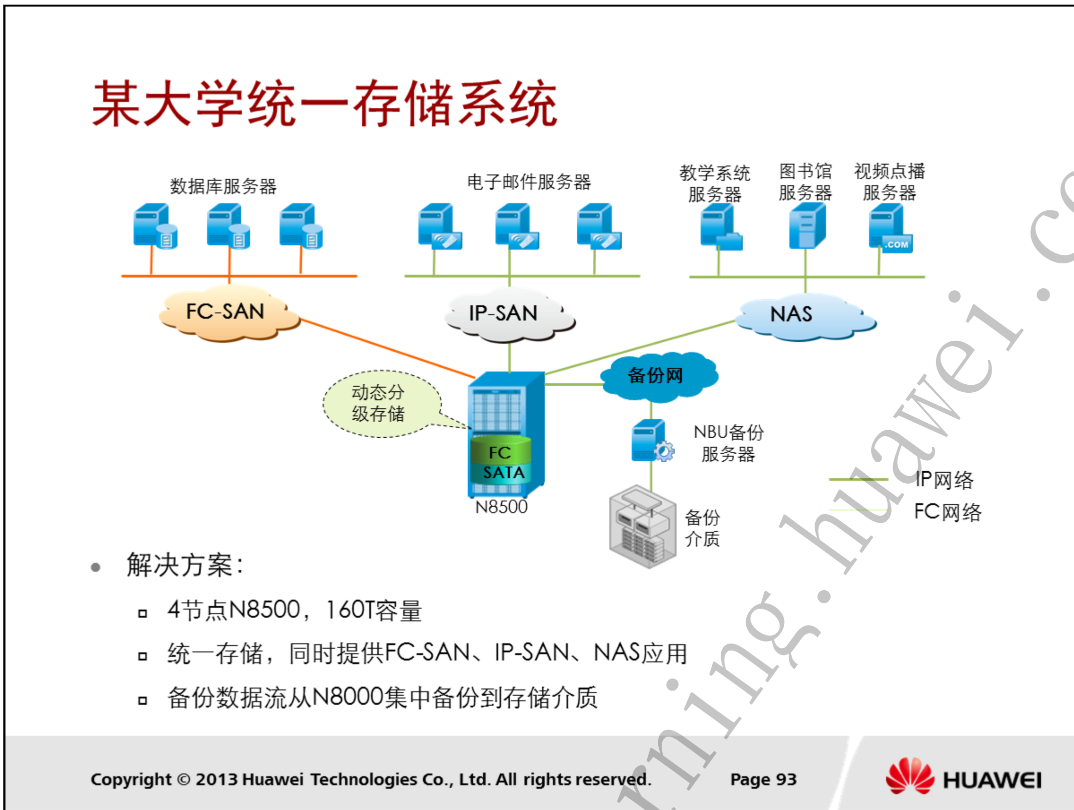
A：N8000的卷镜像功能是同步方式，通俗理解为对文件系统做Raid1，实现镜像保护。这种保护有两种方式：一种是阵列间进行卷镜像（阵列级别的保护，一台阵列故障时，业务不受影响），一种是阵列内进行卷镜像（raid级别的保护，一个Raid组故障时，业务不受影响）。

目录

HUAWEI



- 客户价值：
    - 集中存储，解决全台网高标清制、编、播、存、管等部门分散存储难题
    - 提供高读写带宽，满足30个工作站同时工作
    - NAS引擎并发工作，性能比原系统提高了2倍，满足高清视频的性能需求
    - 数据镜像功能为关键数据提供实时保护
    - 业务持续增长，满足性能和容量的在线扩展需求



N8500对客户的价值包括：

- 通过数据统一存储，降低系统管理、维护、运营成本

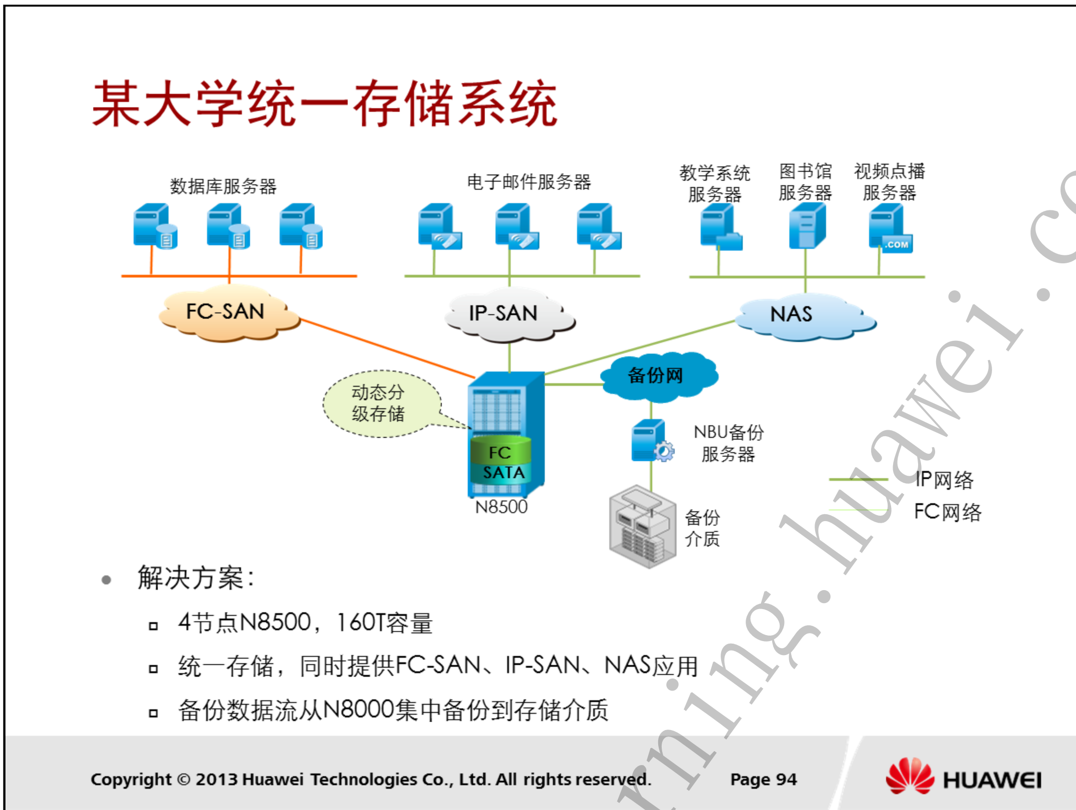
SAN/NAS一体化方案，既能提供块级的数据存储，也能提供文件级数据存储，所有业务系统数据统一存放到N8000存储系统上，实现存储数据的集中存放，集中管理，满足结构化/非结构化数据的存储需求。以前各业务系统独立建设时，既要建立生产系统，还需要建设各自的备份系统，造成资源的浪费。现在采用集中建设生产系统、备份系统，有效的降低系统建设管理、维护成本。

- 很好满足了数字校园各业务系统的SLA（服务等级协议）承诺

一般情况下，学校各部门是系统的使用方，IT中心负责对各业务系统进行运维和管理，应用部门都会对IT部门提出要求，如今年业务系统A宕机时间不能超过一小时，影响范围不能超过500人，关键性数据不能有任何丢失，系统B宕机时间不能超过2小时，可以接受数据少量丢失等。通过N8000集群存储系统对价值数据提供了高可靠的存储、并对关键数据提供备份/容灾保护，兑现了各业务系统的SLA承诺。

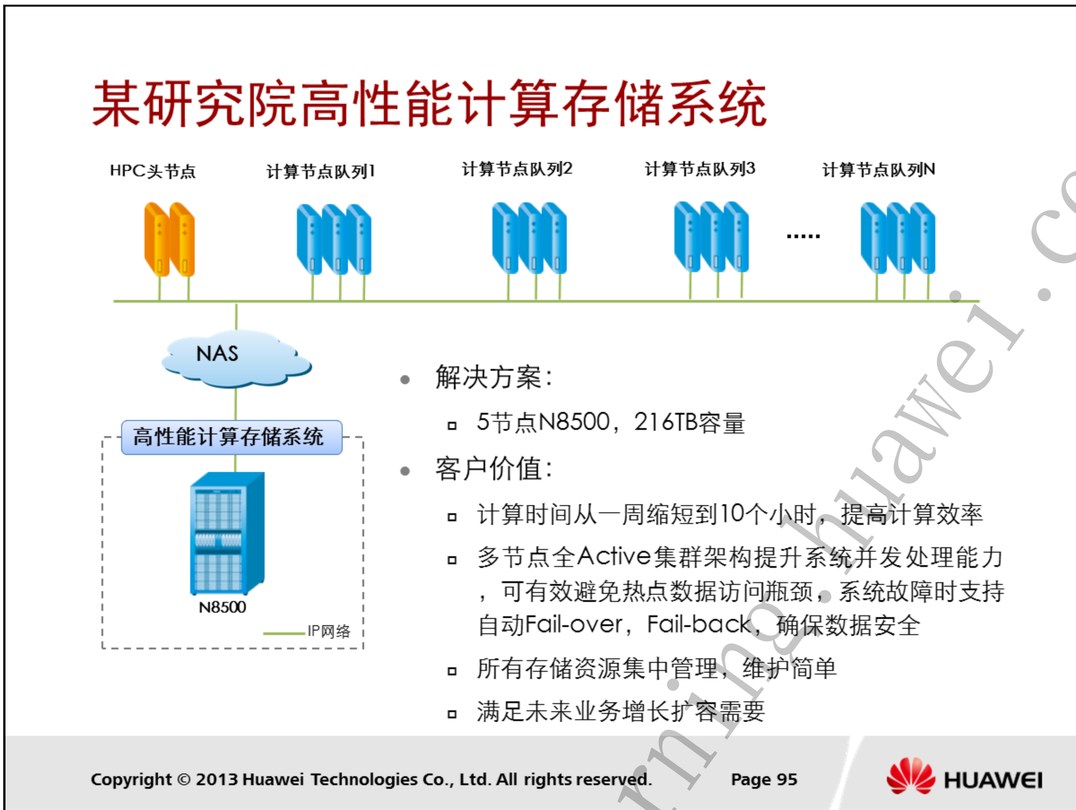
- 通过数据的生命周期管理，降低了系统建设成本

通过DST技术，可以将资源集中用于热点价值数据的处理，降低了系统整体建设成本。按照热点数据和非热点数据2：8比例，同等容量下使用高性能盘的系统建设成本是使用动态分级存储技术系统的建设成本的3-4倍。实现存储资源的动态调配，灵活满足各业务系统动态的存储需求。比如当使用FC-SAN的块存储设备容量不够时，可以将NAS的空闲存储容量调配过去，相反，NAS设备容量不足时，也可将多余的块存储容量调配过来。



- 适应了学校教学和业务未来高速增长的要求

满足未来5～10年，信息量持续增长的扩容的需求，同时满足校园终端和主机客户数量未来持续增长导致的存储性能扩展需求。即性能出现瓶颈的时候，可以横向扩展引擎节点实现性能的近线性增长，存储容量不足时，可以纵向扩展存储单元增加容量。

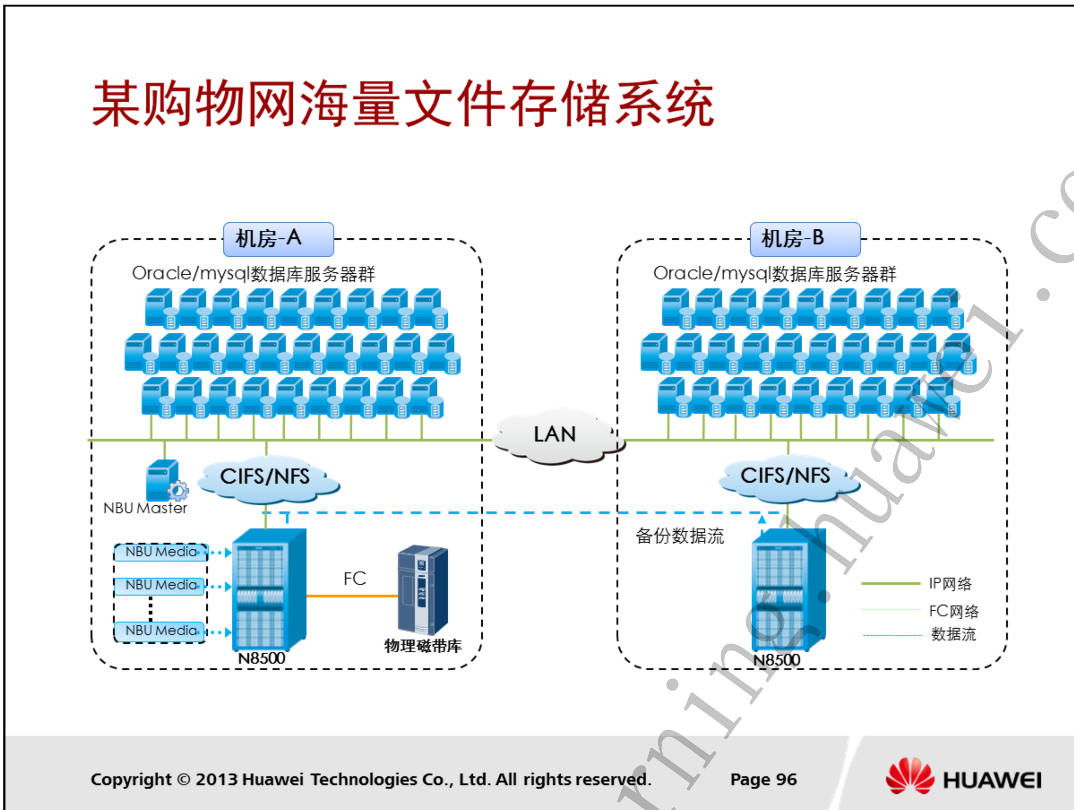


大规模的高性能计算集群对后端存储的要求非常高，如果使用的存储架构是IO节点+FCSAN的形式，其实就是自己搭建的NAS系统。这样搭建系统很简易，只是实现了一个NFS文件系统共享的功能，在性能、可靠性、稳定性、管理维护性方面都难以满足实际的需要，特别是性能不足，导致了高性能集群的效能发挥不出来。

高性能计算对存储的读写基本上属于大文件读写类型，每个计算节点一次读取若干数据，完成计算后，再把结果写回存储，计算节点越多，对存储性能的要求也就越高。之前的IO节点+FCSAN的形式，管理分散，维护复杂，所以集中管理，维护简单也是客户所看重的。高性能计算每天都会生成大量的数据，有时候最多的时候有几百GB的数据，这就要求要有很好的扩容能力。而且随着业务的发展，计算节点的数量也越来越多，这也就要求存储要能够有足够的性能扩展能力。

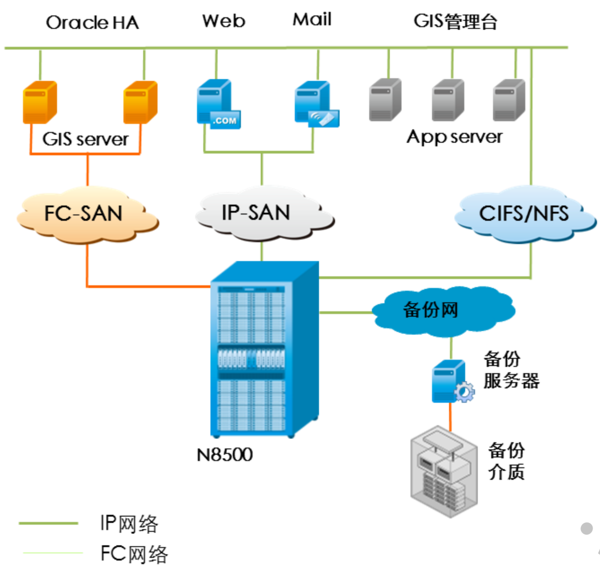
这里以一个计算集群为例，计算节点100台，负责计算任务的执行，可以同时运行多个计算任务，每个计算节点都是整个计算能力资源池中的一个资源，只要资源没有耗尽，就可以执行新的计算任务，如果已经耗尽了，计算任务就需要等待空闲的资源，计算节点和存储系统有大量的交互，要频繁的读取和写入数据。管理节点两台，用于分发任务到各个不同的节点；控制计算的整个流程，管理节点不需要跟后端存储有大量的数据交互，但是也要访问存储空间，因为存储空间上存放着任务执行的脚本。

目前配置了5节点的N8500，三套S5500，每套72块1TB SATA盘，共216T，整套系统实测带宽超过GB/s，充分满足客户需求。



- 解决方案：
  - 内置NBU Media备份软件，提供高带宽数据备份解决方案
- 客户价值：
  - 大容量集中存储系统，支撑客户未来5年爆发性数据增长
    - 提供1496TB大容量存储系统，并可根据客户需求在线扩展至15PB，满足某购物网未来爆发性的数据增长
    - 竞争对手单套系统只能支持不到1PB的存储空间，无法满足用户扩展性需求
  - 备份效率提升4倍，备份更加轻松便捷
    - 12节点并发读写吞吐量近3GB/s，完全满足用户500多台数据库服务器并发访问要求
    - 用户每日数据增长量近15TB，关键数据超过5TB，备份窗口不到4小时
    - 内置NBU Media备份介质软件，实现高性能并发式数据备份，与传统NDMP协议备份方式相比，备份效率提高近4倍

某国土资源局GIS系统



- 解决方案：
  - 4节点N8500
  - 统一存储，将原有FC-SAN，IP-SAN和NAS等多种组网整合

Page 97 HUAWEI

- N8500给客户带来的价值包括：
  - 简化IT基础架构
    - 通过整合，将原有多套单独建设的存储系统整合到一套设备上，简化IT基础架构，降低管理维护成本。
  - 保障业务的连续性和稳定性
    - 对GIS业务系统数据实现高可靠保护，解决原有存储系统存在的单点故障风险，保障GIS系统7×24小时持续运行。
  - 大幅度提升系统整体性能
    - 多个NAS引擎同时提供业务，提升系统性能，解决热点数据性能瓶颈，实现业务负载均衡。
  - 助力业务持续增长
    - 提供在不中断业务的情况下横向添加NAS引擎提升性能，纵向添加存储单元扩展容量的能力，满足未来业务增长需求。

总结

- 集群NAS硬件组成和软件特性
- N8500系统内外组网
- N8500软硬件安装
- NFS和CIFS配置部署
- FS高级特性原理与配置

思考题

- 集群NAS硬件基本组成有哪几部分组成？每部分的主要功能是什么？
- N8000集群NAS系统的IP地址如何规划?
- 请描述一下N8500系统初始化配置和文件系统共享的基本配置的过程?

Thank you

www.huawei.com

HC1209204 大数据存储系统部署与管理

HC1209204 大数据存储系统部署与管理

www.huawei.com

HUAWEI

目标

- 学完本课程后，您将能够:
  - 描述大数据存储硬件组成和软件特性
  - 掌握9000系统典型组网
  - 掌握9000软硬件安装
  - 熟悉NFS和CIFS配置部署
  - 熟悉Info系列特性原理与配置

 Page 1 HUAWEI

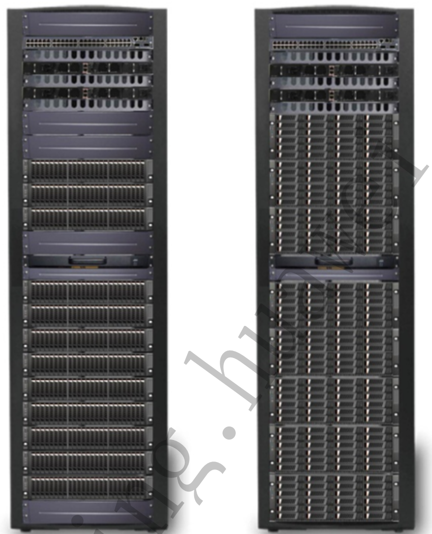
- 学习完本章具备OceanStor9000大数据存储产品部署与管理能力。

目录

HUAWEI

大数据存储OceanStor9000介绍

- 大数据存储特点
  - 弹性空间
  - 卓越性能
  - 简化管理
  - 高可靠性
- 应用场景
  - 广电媒资
  - 高性能计算
  - 大型企业数据中心
  - 互联网运营



P节点机柜　　C节点机柜

Page 3

HUAWEI

卓越性能

- 内部高速互联：采用10GE Ethernet或40GE Infiniband网络互联，支持TOE/RDMA技术
- SSD元数据加速：采用SSD盘存储元数据，提高元数据访问效率
- 全局缓存：最大可支持55TB全局缓存，提高数据访问命中率
- 动态分级存储：热点数据智能迁移，提升系统性能
- 智能负载均衡：跨节点的客户端连接负载均衡，自动平衡容量和性能，优化集群资源

弹性空间

- 弹性扩展：超乎想象的3-288节点无缝扩展，容量、性能随节点增加而线性增长
- Scale-out架构：可灵活扩展的Share-nothing的全对称分布式架构
- 全局命名空间：最大40PB的单一文件系统

数据可靠

- 数据冗余保护：提供业界最高水平的数据可靠性和可用性；系统全冗余部署，无单点故障
- 可靠性灵活配置：支持基于目录的冗余配比策略，提供不同的数据保护级别
- 自动修复：智能逻辑坏道、物理坏道修复，并采用磁盘预拷贝技术，降低磁盘故障率
- 权限管理：提供针对IP地址或用户/用户组的访问权限控制，创建安全、隔离的存储池

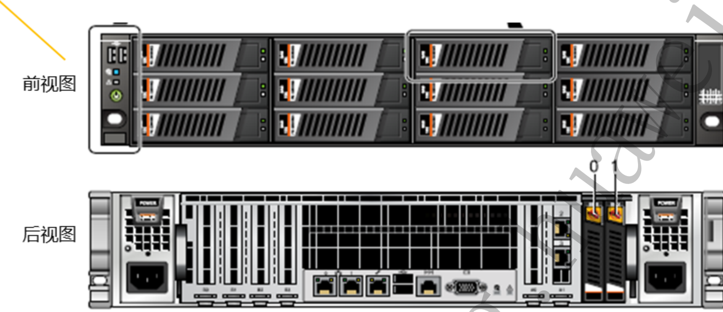
应用场景

- OceanStor9000作为大数据存储产品，可广泛应用于广电媒资、高性能计算、数据中心、互联网运营和大型企业等各行业领域。

OceanStor 9000硬件节点



Mini-Capacity node——小容量应用场景



2U设备

12个数据硬盘槽位（3.5“盘）

标配：2路4核CPU，32G内存，

12*2TB SATA盘（数据盘）

 Page 4 HUAWEI

OceanStor 9000家族硬件



Performance node——OPS密集型应用场景

前视图



后视图



2U设备

25个硬盘槽位（2.5“盘）

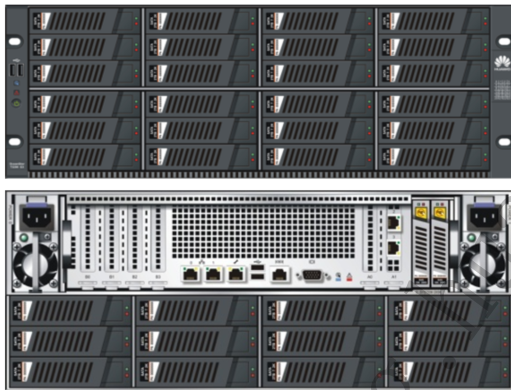
标配：2路6核CPU，48G内存，

支持SSD、SAS盘，按需配置

OceanStor 9000家族硬件



Capacity node——大容量应用场景



4U设备

36个硬盘槽位（3.5“盘）

标配：2路6核CPU，48G内存

支持SSD、NL SAS、SATA盘，按需配置

目录

HUAWEI

硬件组件



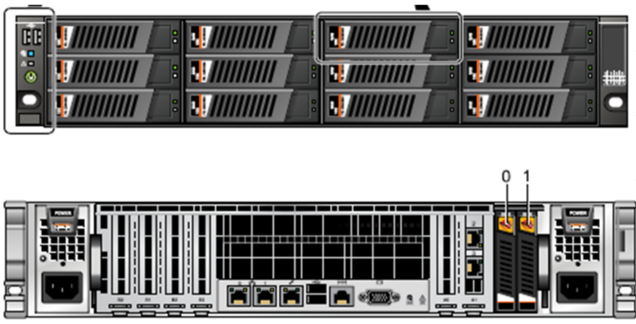
类型	可选型号	说明
存储节点	M节点	2U存储节点
	P节点	2U存储节点
	C节点	4U存储节点
网络设备	CE6850-48S4Q-EI	10GE以太网交换机
	S5700-52C-SI	GE以太网交换机
	S5352C-SI	GE以太网交换机
	IS5023	IB（InfiniBand）交换机
8口KVM控制器		提供8路键盘、鼠标和视频端口

 Page 8 HUAWEI

- 主要包括:
    - 存储节点（P和C节点），负责承载分布式存储系统软件和各种特性，同时提供存储空间资源
    - 连接设备：主要负责管理网络、后端集群网络和前端业务网络互联，采用高速互联网络主要解决大数据存储、访问的高带宽需求

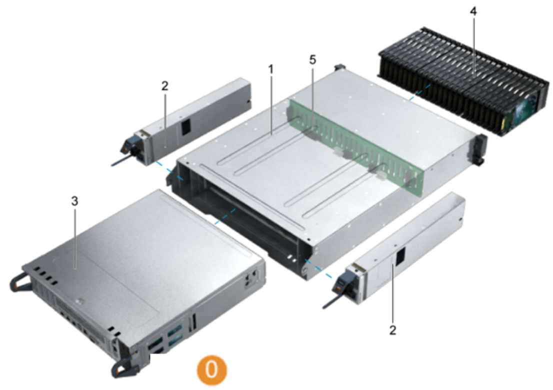
M节点硬件形态

- Mini-Capacity node，支持小容量应用场景



高度	2U
最大功耗	526.6W
CPU	2×Intel E5 2403 4核 1.8GHz
内存	24GB~96GB，32GB
操作系统硬盘	2×300GB 10K rpm 2.5寸 SAS (RAID1)

P节点形态



1	机框	2	电源模块
3	主控板插框	4	业务硬盘模块
5	业务硬盘背板	-	-



Page 10

HUAWEI

- P节点前面板提供25个业务硬盘槽位，作为系统数据盘，从左到右依次编号为0～24
- 0号～3号槽位支持4个2.5英寸的业务硬盘（SSD），作为系统数据盘，用于提供元数据存储和分集存储特性的高性能存储空间
- 4号～24号槽位支持21个2.5英寸的业务硬盘（SAS），作为系统数据盘，用于提存储系统的物理空间

P节点端口

4 PCI-E扩展插槽

5 SAS卡

6 GE网口

7 IPMI管理网口

8 USB端口

9 系统串口

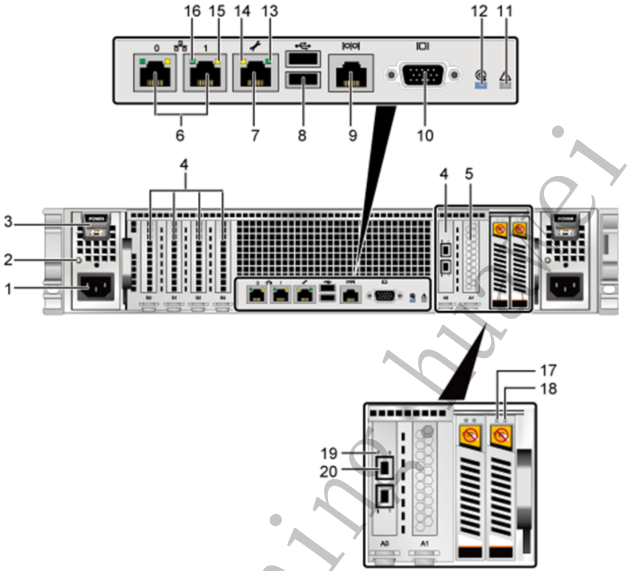
10 VGA端口

11 系统告警指示灯

12 系统定位指示灯

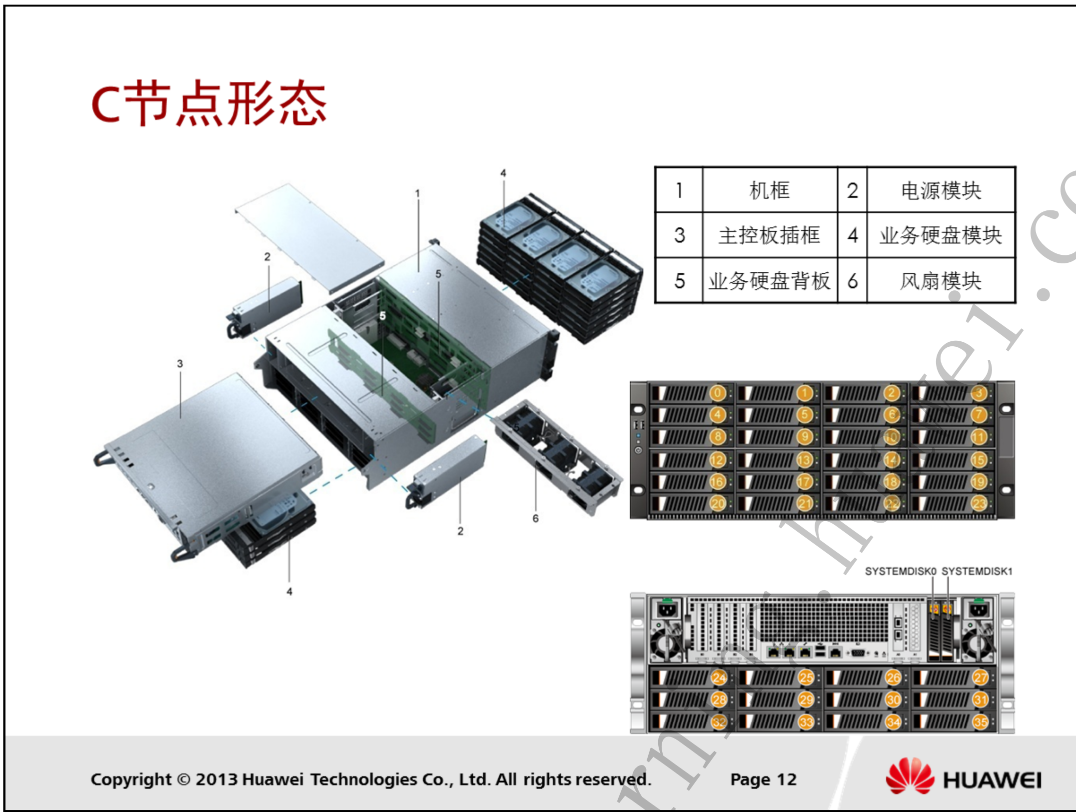
18 硬盘告警/定位指示灯

20 10GE SFP+以太网光接口

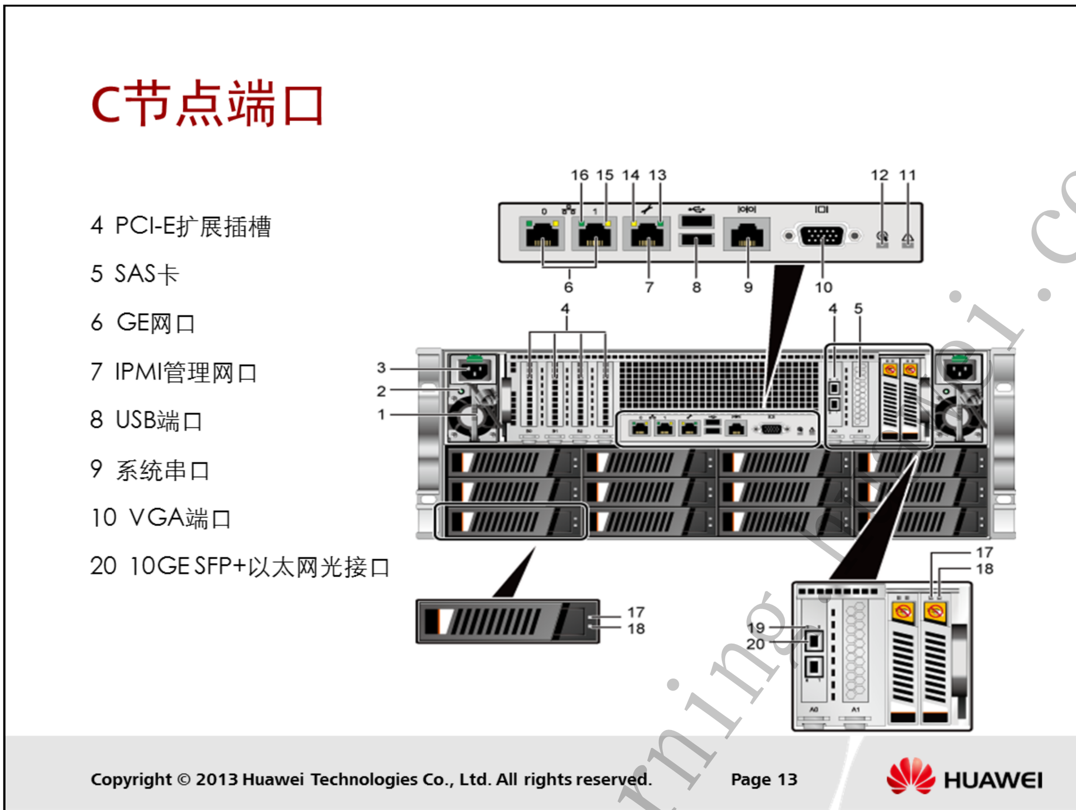


 Page 11 HUAWEI

- P节点后面板提供2个操作系统硬盘槽位，从左到右依次编号为SYSTEMDISK0和SYSTEMDISK1,为2块300GB 2.5英寸SAS硬盘，用于承载节点本地操作系统
- P节点的PCI-E扩展插槽可选配：2端口GE网卡，2端口TOE 10GE网卡，2端口IB网卡或SAS RAID卡



- C节点前面板提供24个业务硬盘槽位，从左到右，从上到下，依次编号为0～23，左边最顶端的槽位是0号槽位
- C节点后面板提供12个业务硬盘槽位，业务硬盘槽位编号延续前面板的序号，从左到右，从上到下，依次编号为24～35，右边最底端的槽位是35号槽位
- 数据盘：1个3.5英寸的业务硬盘200G SSD，插入0号槽位 ；35个3.5英寸的业务硬盘4TB SATA，插入1号～35号槽位
- 系统盘：支持2个2.5英寸的后置操作系统硬盘300GB SAS



- C节点后面板提供2个操作系统硬盘槽位和12个业务硬盘槽位， 系统盘槽位从左到右依次编号为SYSTEMDISK0和SYSTEMDISK1
- C节点的PCI-E扩展插槽可选配：2端口GE网卡，2端口TOE10GE网卡，2端口IB网卡或SAS RAID卡

附：硬件特性参考

型号	P节点	C节点
系统架构	全对称分布式架构	
节点数	3～288	
CPU/节点	2×Intel E5系列	
缓存/节点	标配48GB，可扩展至192GB	
硬盘类型	数据盘支持2.5寸SSD、SAS硬盘（根据客户实际性能需求，可调整SSD/HD的配比）	数据盘支持3.5寸SSD、SATA、NL SAS硬盘（根据客户实际性能需求，可调整SSD/HD的配比）
硬盘数/节点	标配4×2.5" 200G SSD + 21×2.5" 600G SAS	标配1×3.5" 200G SSD + 35×3.5" 4TB SATA
前端网络	10GE Ethernet或40GE Infiniband或1GE Ethernet	
内部互联	10GE Ethernet或40GE Infiniband或1GE Ethernet	
应用场景	OPS密集型应用场景	高带宽应用场景

Page 14

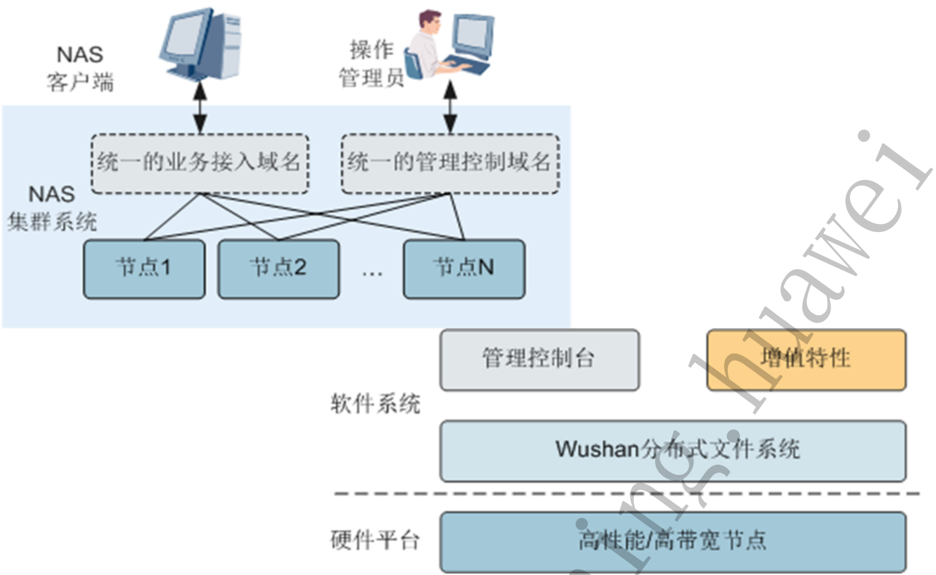
HUAWEI

- 此特性是V100R001C00版本特性，产品特性随产品版本变化
- 重点了解：硬盘类型、前端网络、内部互联

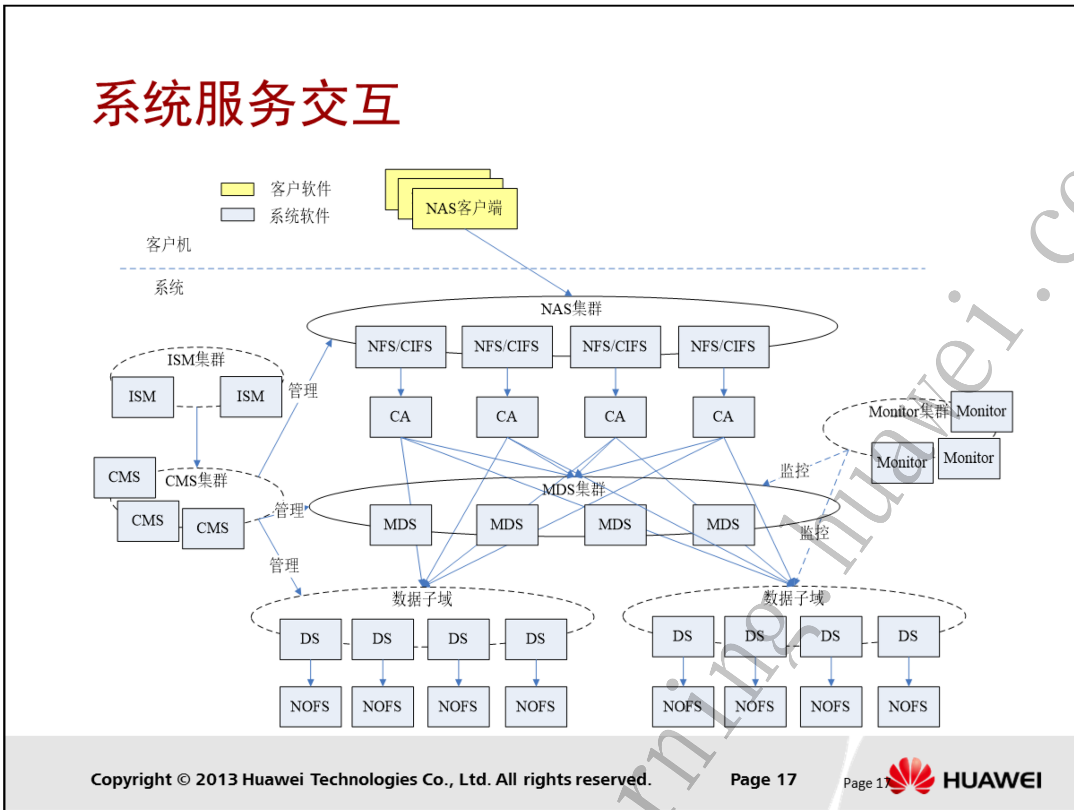
目录

HUAWEI

9000系统组成



HUAWEI



- NFS/CIF集群负责NAS协议；
- CA（客户端代理）负责文件系统协议的语义解析执行，是文件系统业务发动机；
- MDS（元数据服务）管理文件系统的元数据，可以定义为文件系统元数据的读写缓存；MDS集群不同节点按照子树切割；相互不重叠；考虑负载均衡，扩展性，可靠性，提供子树迁移能力
- DS（数据子域）集群提供可靠的对象存储；为文件系统元数据和文件数据提供存储能力；DS集群提供差异化的数据冗余保护能力；
- Monitor集群提供系统的集群状态监控和状态表格同步，可靠保存的能力；
- CMS（集群管理服务）集群对集群的物理节点进行设备管理，提供配置、告警、日志、升级能力；
- ISM为OceanStor 9000提供配置，告警等UI界面

Info系列软件



WushanFS：Wushan File System

1 WushanFS

- Wushan FS：分布式文件系统
- Info Data Protection 高可靠数据保护

2 InfoLink

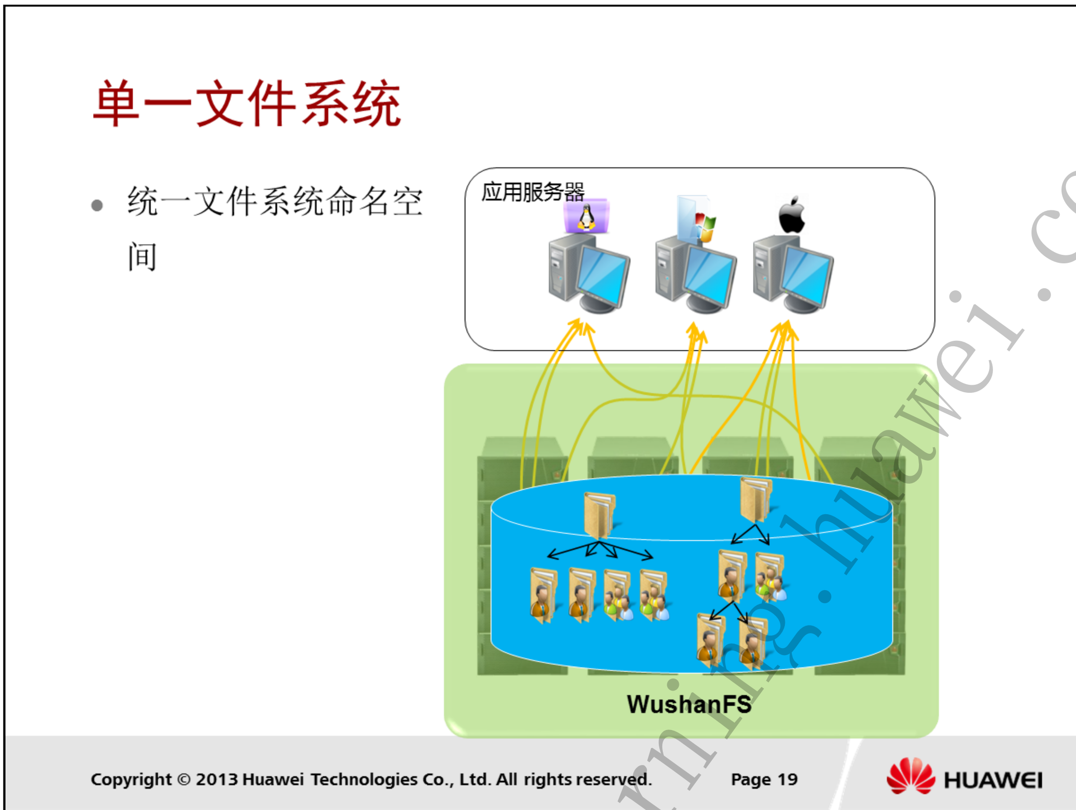
- 集群负载均衡软件
- 多种负载策略支持
- 可选软件，单独购买license

3 InfoQuota

- 配额管理,可对目录或用户进行空间限制
- 可选软件，单独购买license

4 InfoTier

- 自动分级存储
- 基于策略的自动迁移
- 可选软件，单独购买license

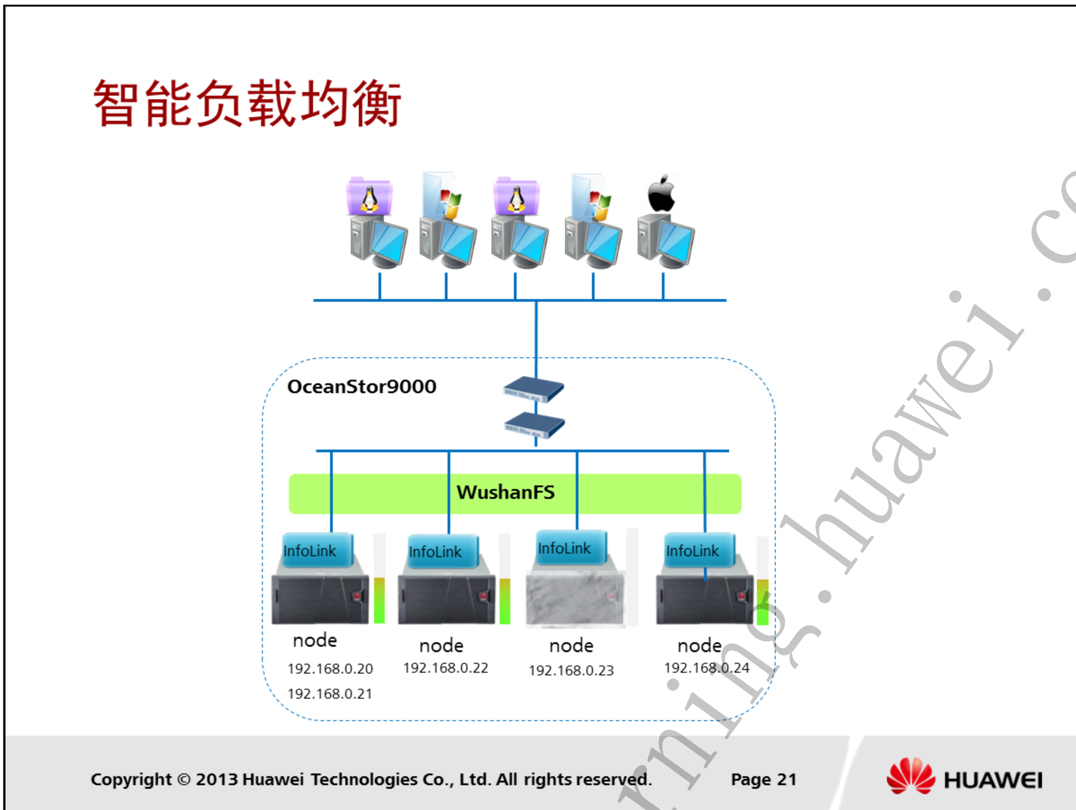


OceanStor 9000对外提供统一的文件系统命名空间。用户存储文件、设置共享、管理权限的操作对象均位于该命名空间内。统一的命名空间可以使用和管理当前系统的全部可用容量。

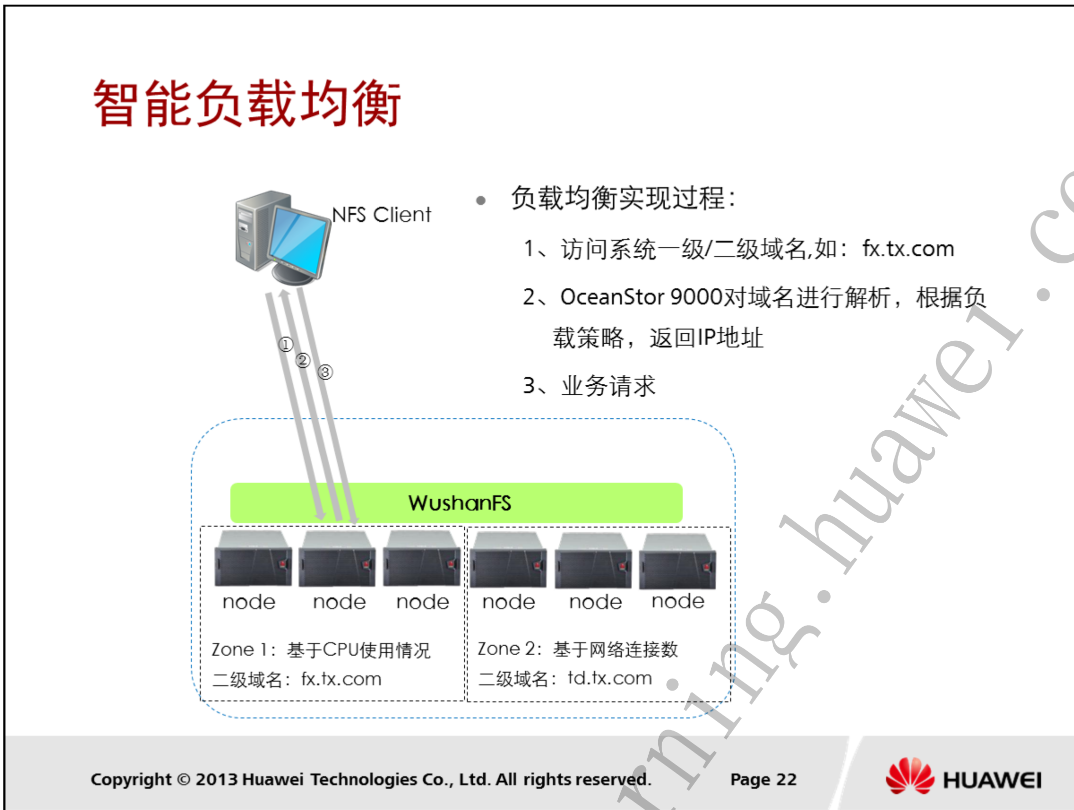
命名空间在OceanStor 9000启动时自动创建，其名称与系统命名保持一致。各共享目录均为命名空间的子目录。同时命名空间作为整个系统的根目录，也可以被共享和访问。

统一的命名空间为用户提供统一的存储资源池，实现对存储空间的集中管理和分配。当进行扩容时，新增容量可直接纳入资源池使用，避免维护多个文件系统。

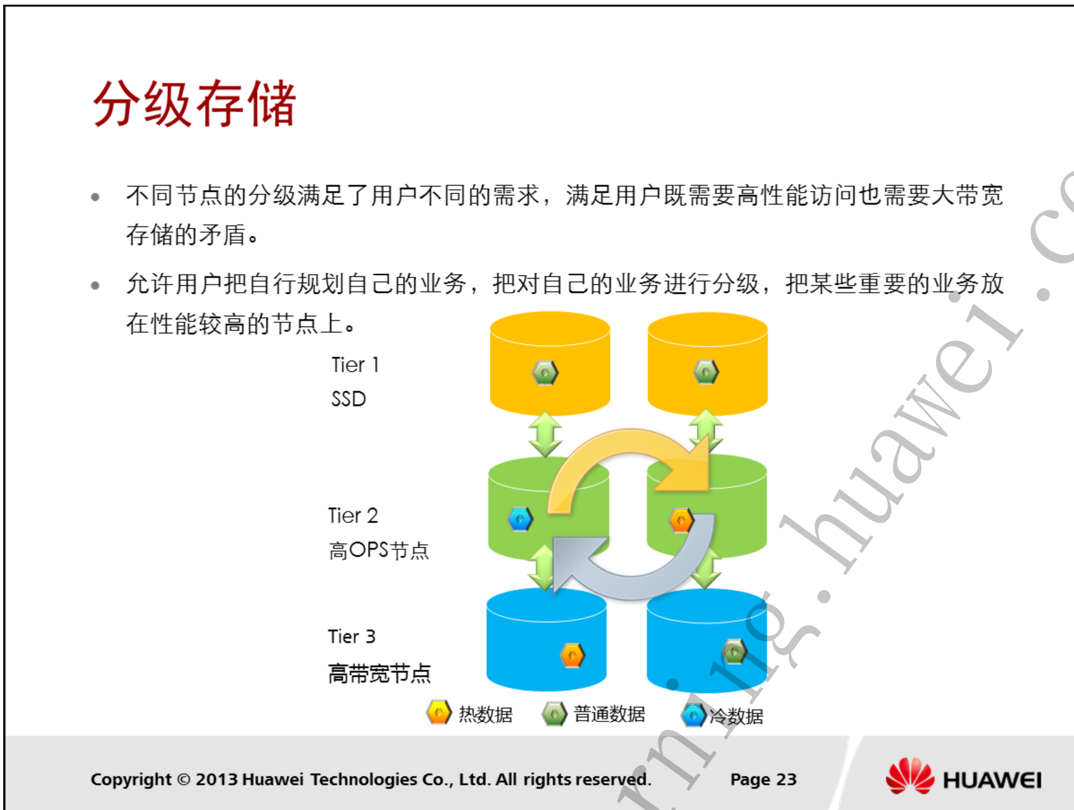

按需扩展
应用服务器
3-288节点线性扩展
60秒之内完成单节点扩容
无需人为干预，无需任何改动
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 20
HUAWEI



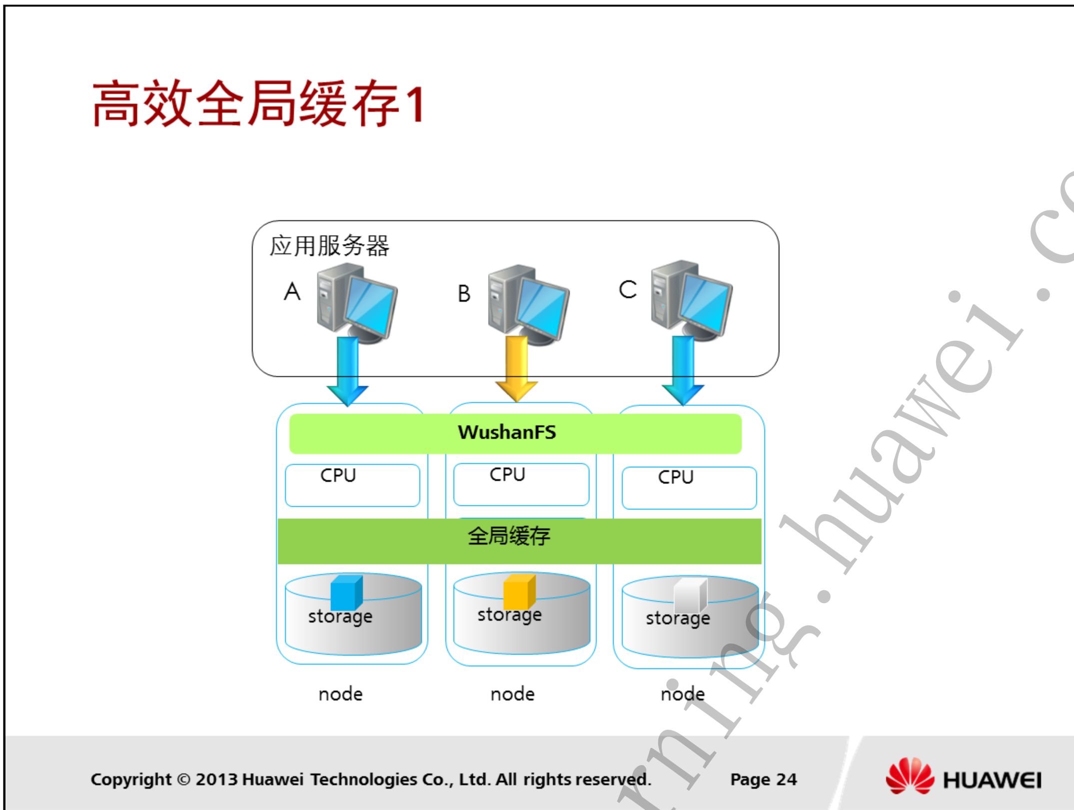
- 在所有节点间自动进行负载均衡
- 自动平衡容量和性能
- IP的故障切换(failover)和故障恢复(failback)
- 无需人工干预，无需更改配置
- 对应用透明



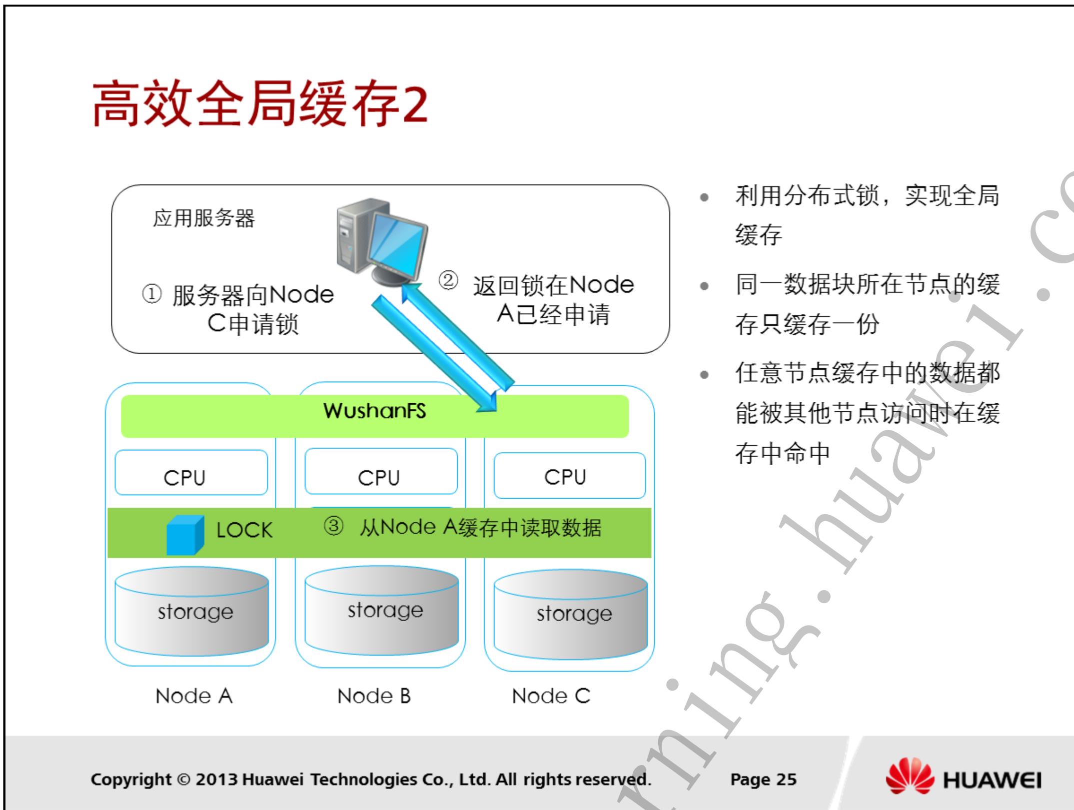
- 智能负载均衡特点:
  - 对外提供单一域名访问方式
  - 可划分分区，对不同分区设置二级域名，以及不同的负载策略
  - 负载均衡策略可以基于：CPU、内存、带宽、网络连接数、节点能力值



- 分级存储特点：
  - 支持不同类型节点间分级存储
  - 基于自定义的分级策略
  - 支持文件大小，路径，时间属性，IO热度等策略
  - 无缝地适应业务变更，而不会影响应用程序
  - 将多个应用程序整合到单个存储系统中，消除了数据移动且简化了管理



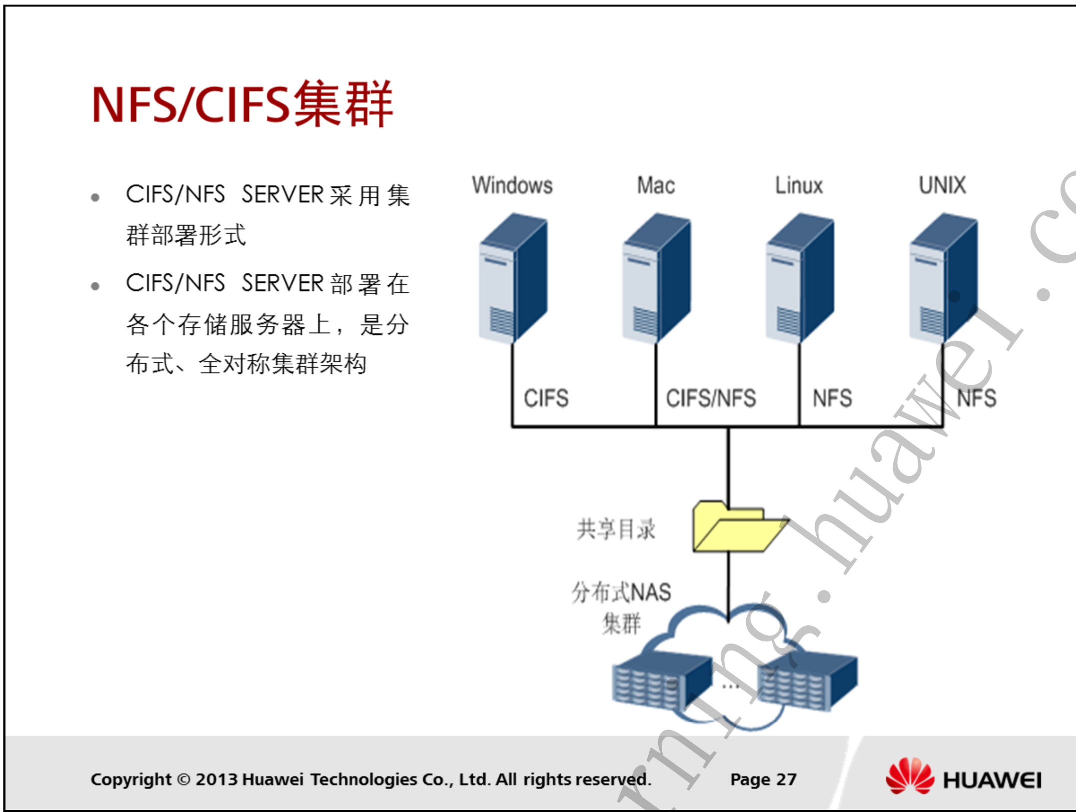
- 高效全局缓存特点：
  - 将集群中所有节点的缓存整合成统一的全局可见统一的大缓存
  - 文件条带数据的全局缓存且只缓存用户数据，不缓存冗余数据
  - 缓存热点动态迁移
  - 55TB超大缓存池



- 如果其它客户端持有该锁时，需要从其它客户端缓存中读取数据
- 如果没有客户端持有该锁，需要申请锁，重建条带数据



- 提供基于目录或用户/用户组的配额控制，并可以在存储达到限制时发出告警。
- 提供两种类型的配额
  - 计算配额：只监控容量使用情况，不限制使用
  - 强制配额：在监控容量使用情况的同时，可以控制硬盘使用
- 配额设置限制：整系统中最大支持8192个配额项，配额不限制文件数，对windows和linux区分配额数



CIFS/NFS共享的单位是目录。所共享的目录可被多个客户端访问。OceanStor9000各节点以集群方式对外提供共享服务，节点间可相互监控服务状态。

OceanStor9000可根据业务和节点状态实现负载均衡，数据访问在集群内均匀分布，以保证高并发场景下的系统性能。当出现设备故障后，OceanStor9000可自动切换服务节点，以保证业务的高可靠性。

附：软件特性参考

型号	P节点	C节点
数据保护级别	N+1，N+2，N+3，N+4	
文件系统	Wushan分布式文件系统支持全局命名空间，可动态扩展，最大40PB	
增值特性	动态分级存储 (InfoTier)/客户端连接自动负载均衡 (InfoLink)/空间配额管理 (InfoQuota)	
自动精简配置	支持自动精简配置，无需配置	
数据自愈	数据自动化并行快速恢复，最高恢复速度达1TB/Hr	
系统扩展	“一键式”在线扩展，单节点扩展小于60秒	
全局缓存	支持全局缓存，最大可支持55TB	
支持的操作系统	Windows，Linux，Mac OS	
协议支持	NFS，CIFS，NIS，Microsoft Active Directory，LDAP，SNMP	
免及时维护	系统自动坏盘率检测及告警通知，坏盘可批量集中更换，无须即时更换，减少人力维护	

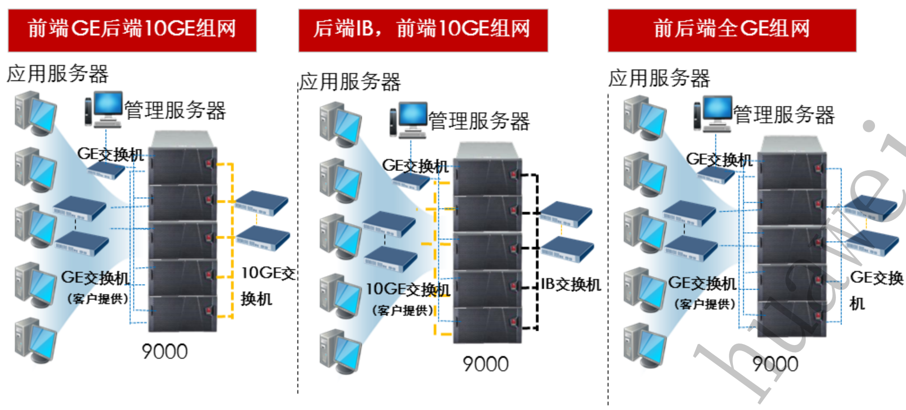
 Page 28 HUAWEI

此特性是V100R001C00版本特性，产品特性随产品版本变化。

目录

HUAWEI

9000典型组网方式

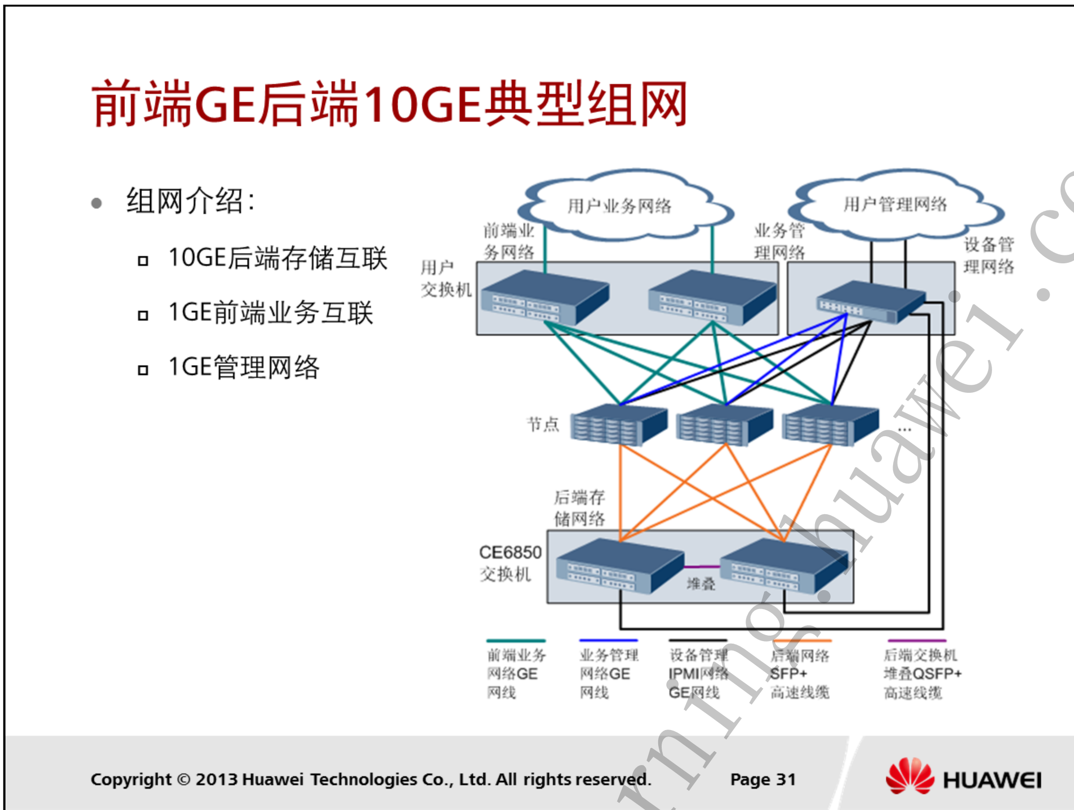


后端网络	前端网络	
10GE	1GE	推荐组网
IB	10GE	
GE	GE	

10GE
IB
1GE

 Page 30 HUAWEI

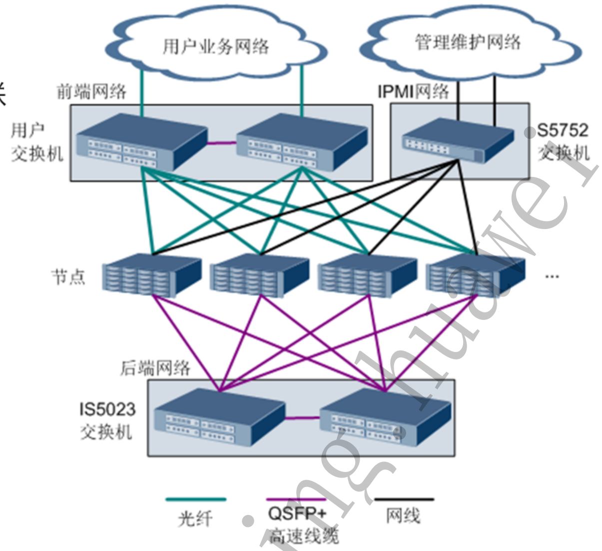
- OceanStor 9000的组网结构包括：
  - 前端业务网络用于 9000与用户网络对接。
  - 后端存储网络用于9000内部节点间互联。
  - IPMI网络用于 9000与用户维护网络对接。
  - 9000支持多种组网方式，包括10GE组网、IB组网和GE组网等，可满足用户的不同组网需求。



- 组网介绍:
  - 各节点使用2个10GE接口，通过高速线缆分别连接到2台CE6850交换机。
  - CE6850交换机之间使用2×40GE QSFP+高速线缆完成堆叠连接。
  - 各节点的IPMI接口使用1×GE网线连接到S5752交换机。S5752交换机通过2×GE网线上行连接用户维护网络。

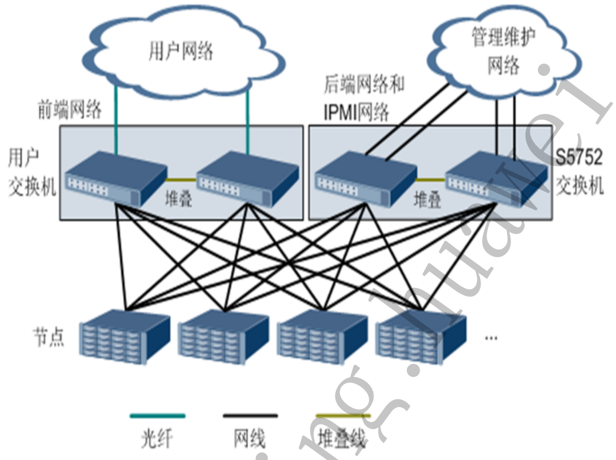
IB组网

- 组网介绍：
  - 40GE IB 后端存储互联
  - 10GE前端业务互联
  - 1GE管理网络

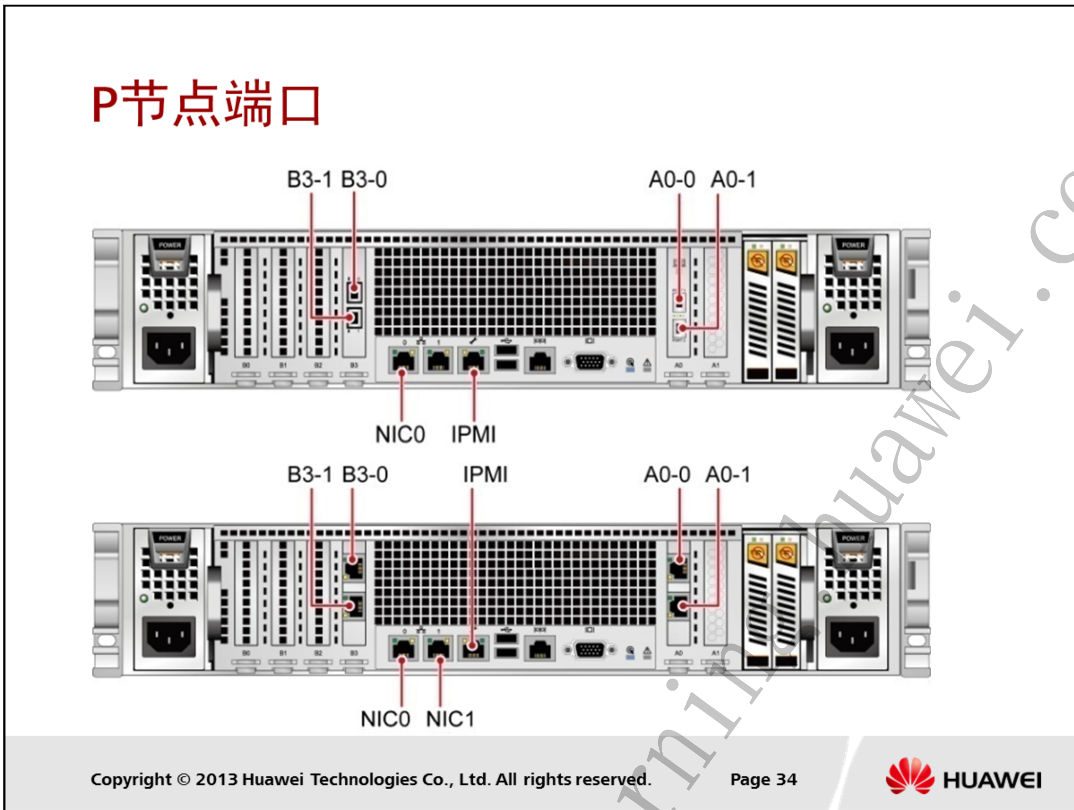


GE组网

- 组网介绍：
  - 1GE后端存储互联
  - 1GE前端业务互联
  - 1GE管理网络



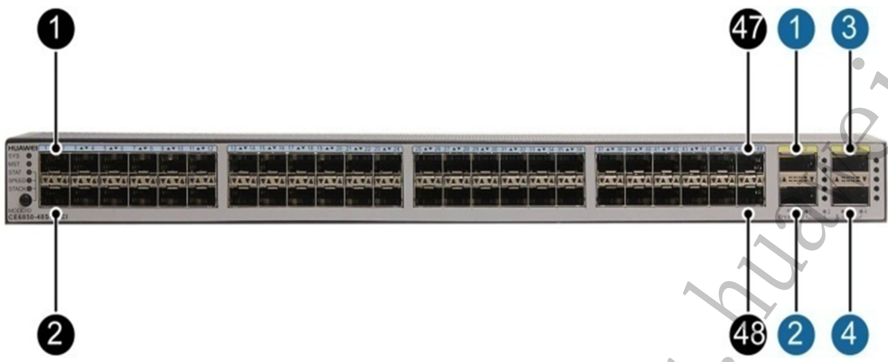
 Page 33 HUAWEI



- 组网方式：前端网络采用客户网络的10GE交换机，后端网络采用IB交换机组网
  - A0-0和A0-1：后端网络，使用2根QSFP+高速线缆连接到IB交换机IS5023
  - B3-0和B3-1：前端网络，使用2根光纤连接到客户网络的10GE交换机
  - IPMI：维护网络，通过1根网线连接到S5752/S5352交换机
  - NIC0：维护网络，将基础柜（9000基础配置机柜）内最下方3个节点的NIC0分别通过1根网线连接到S5752/S5352交换机
- 组网方式：前端和后端网络均采用GE交换机组网
  - A0-0、A0-1、B3-0和B3-1：后端网络，使用4根网线连接到所在机柜中的S5752/S5352交换机
  - NIC0和NIC1：前端网络，通过2根网线连接到用户网络的GE交换机。
  - IPMI：机柜内部编号为奇数的存储节点顺序连接到S5752/S5352-01，编号为偶数的存储节点顺序连接到S5752/S5352-02。

10GE互联组件

- 10GE交换机CE6850-48S4Q-EI

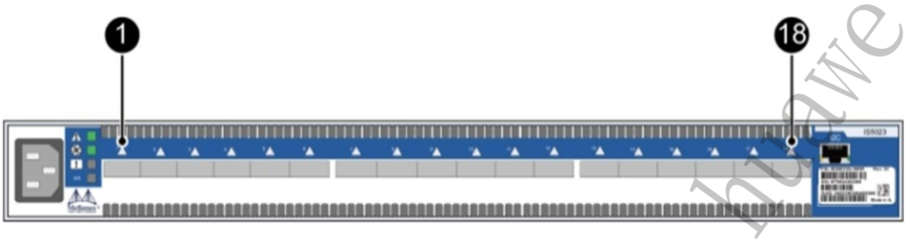


 Page 35 HUAWEI

- 1-48为10GE SFP+以太网光接口，其中：
  - 通常1～32号光接口下行连接到存储节点
  - 33～48号光接口预留
- 右边1-4端口为40GE QSFP+以太网光接口，用于机柜内两台交换机之间堆叠，其中：
  - 节点数量≤16时，使用2个接口堆叠
  - 16＜节点数量≤24时，使用3个接口堆叠
  - 24＜节点数量≤32时，使用4个接口堆叠

40G互联组件

- IB交换机IS5023



- 1~18为40GE QSFP+以太网光接口，其中：
  - 通常1～16号接口下行连接到存储节点
  - 通常17和18号接口为交换机之间的互连接口

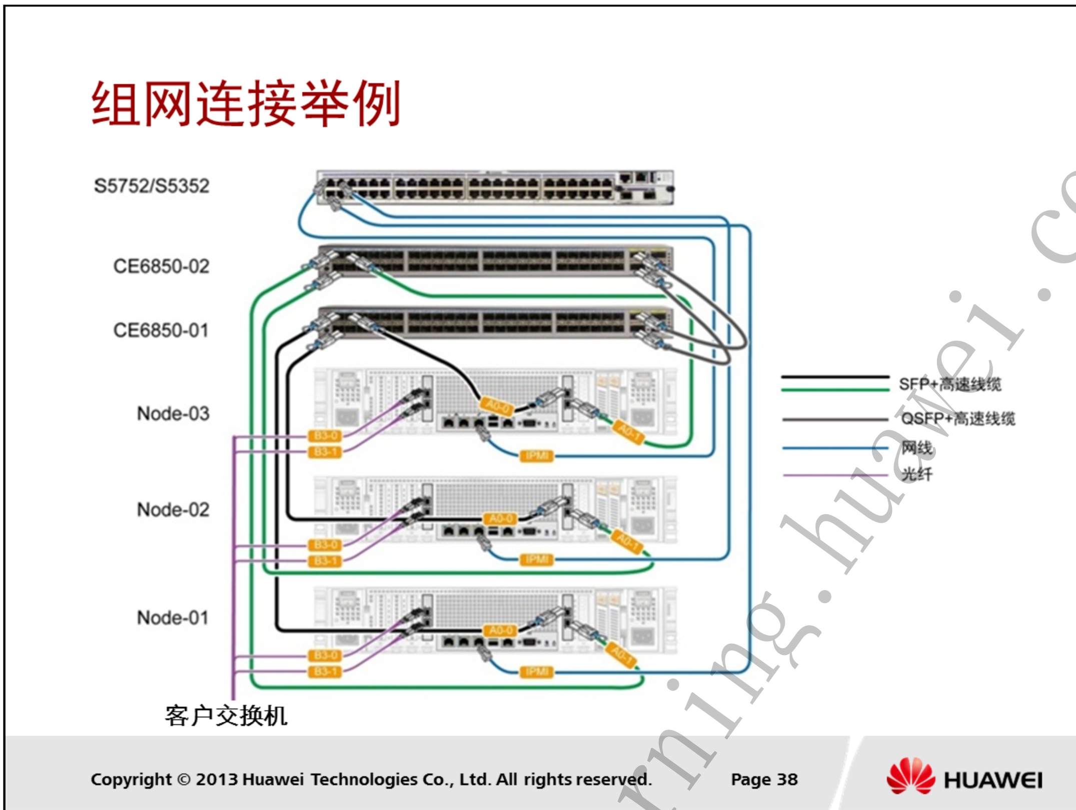
1GE互联组件

- GE交换机S5700-52C-SI/S5352C-SI



 Page 37 HUAWEI

- 10/100/1000BASE-T以太网接口，其中：
  - 在前端网络和后端网络均采用GE交换机时，1～24号接口下行连接到存储节点的GE电接口；25～30号接口下行连接到存储节点的IPMI口；31和32号接口上行连接到客户的网络设备
  - 在前端网络和后端网络均采用10GE交换机时，1～32号接口下行连接到存储节点的IPMI口，42、43和44号接口下行连接到基础柜内1～3号存储节点的NIC0口，45和46号接口连接到10GE交换机的ETH接口，47和48号接口上行连接到客户的网络设备
  - 在后端网络采用IB交换机时，1～32号接口下行连接到存储节点的IPMI口，44、45和46号接口下行连接到基础柜内1～3号存储节点的NIC0口，47和48号接口上行连接到客户的网络设备



- 本节示例中，组网方式为前端和后端网络均采用10GE交换机组网， 部署3台P节点。

IP地址规划1

段类型	网段属性	组网类型[a]	IP地址数目	建议
前端业务网段	为向存储节点存取文件的网络	10GE组网	节点数量×6+1	1. 2个接口IP地址 2. 4个预留作为InfoLink的动态前端IP地址 3. InfoLink DNS IP地址1个 （不论是否购买InfoLink特性License均需要）
		IB组网	节点数量×6+1	
		GE组网	节点数量×6+2	1. 2个接口IP地址 2. 4个预留作为InfoLink的动态前端IP地址 3. InfoLink DNS IP地址1个 （不论是否购买InfoLink特性License均需要） 4. 管理IP地址1个

a：OceanStor 9000的组网类型以后端存储网络交换机的类型区分，如后端存储网络交换机采用10GE交换机，则称此OceanStor 9000组网为10GE组网。

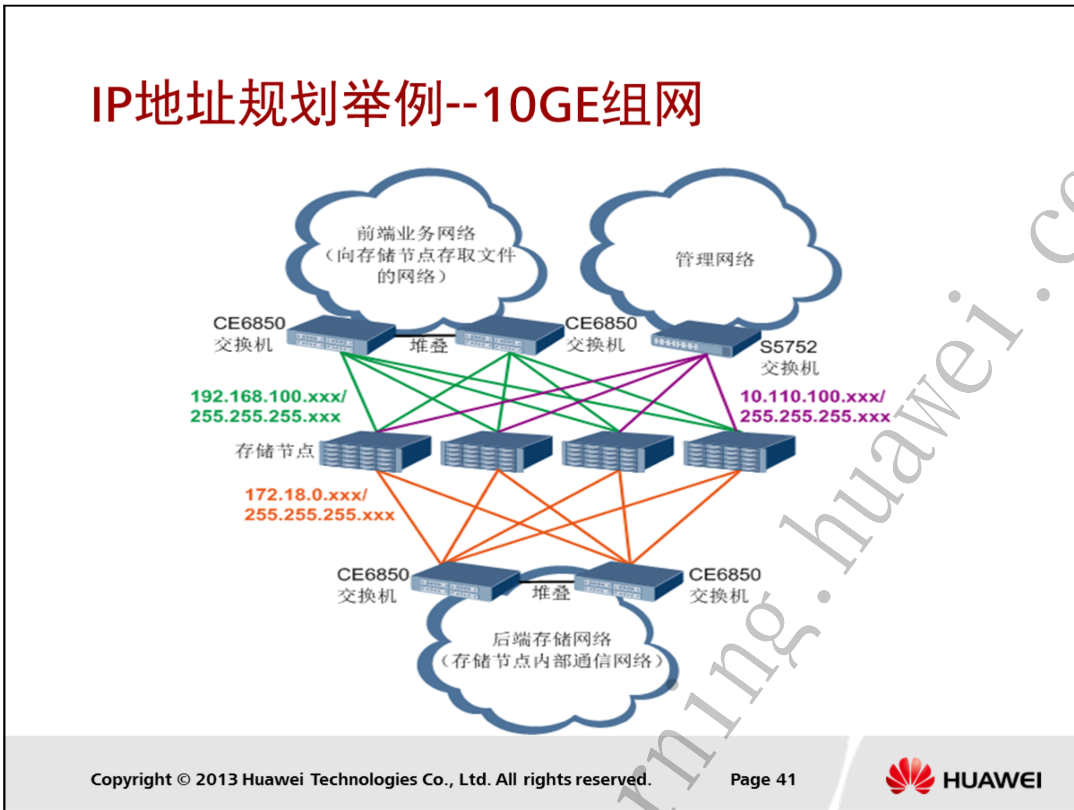
HUAWEI

- 如果后续存在扩容需求，为避免扩容时重新调整网段，请提前为扩容节点和交换机预留IP地址。
- 建议每个网段预留2个以上IP地址，供调试时使用。

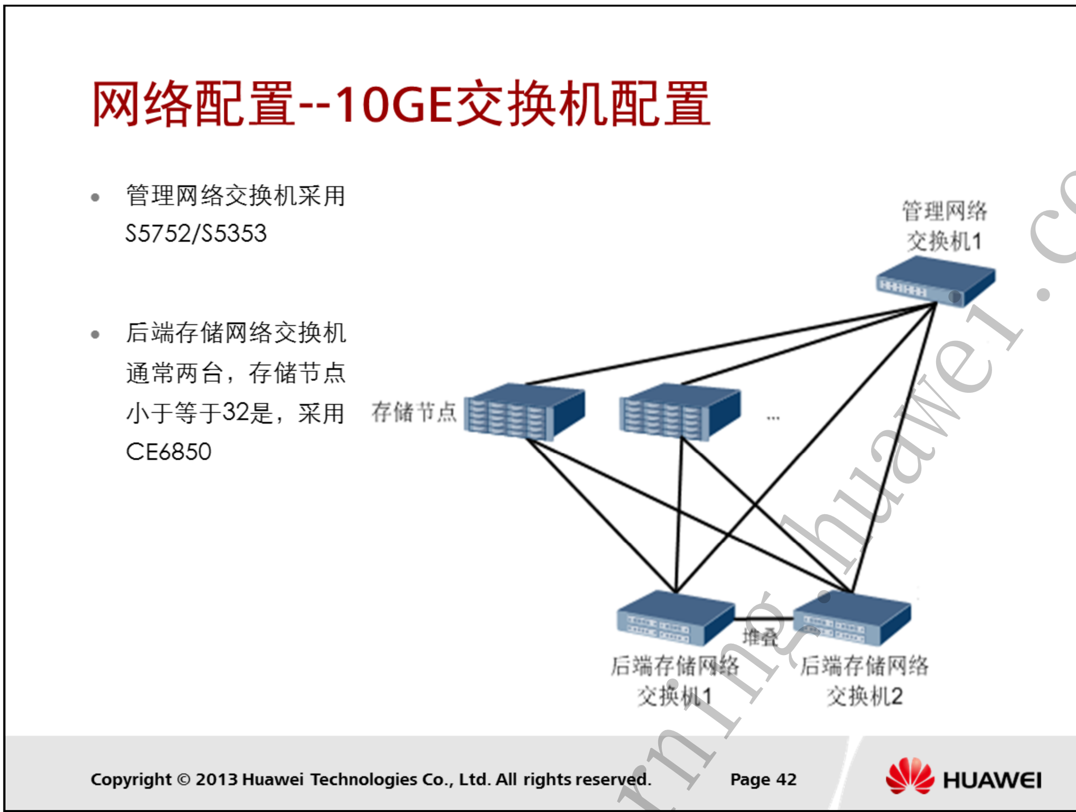
IP地址规划2

网段类型	网段属性	组网类型a	IP地址数目	备注
管理网段	系统管理控制网络	10GE组网	节点数量+管理网络交换机数量+2	存储节点IPMI接口分配1个每台管理网络交换机分配1个，OceanStor 9000管理IP地址1个，后端存储网络交换机1个
		IB组网	节点数量+管理网络交换机数量+1	
		GE组网	节点数量+后端存储网络与管理网络交换机总数量/2	每个存储节点IPMI接口分配1个，每2台后端存储网络与管理网络交换机分配1个
后端存储网段	存储节点间通信网络	10GE组网	节点数量×3	每个存储节点分配2-3个，软件安装时通过DHCP分配临时IP地址时每个存储节点再分配1个。
		IB组网	节点数量×3	
		GE组网	节点数量×2	

HUAWEI



- 前端业务网段：需要的IP地址数目为25，因此掩码长度要求小于或等于27。此例中使用27位掩码。
- 管理网段：需要的IP地址数目为7（未含预留数目），因此掩码长度要求小于或等于28。此例中使用28位掩码。
- 后端存储网段：需要的IP地址数目为12，因此掩码长度要求小于或等于28。此例中使用24位掩码。



下面以两台CE6850、一台S5752为例说明交换机的配置方法。其他型号的交换机配置方法与此类似，您可以参考随交换机发货的相关文档完成对应操作。

CE6850-1超级终端通信参数为9600bps、8位数据位、1位停止位、无奇偶校验和无数据流控制。

- 配置流程：

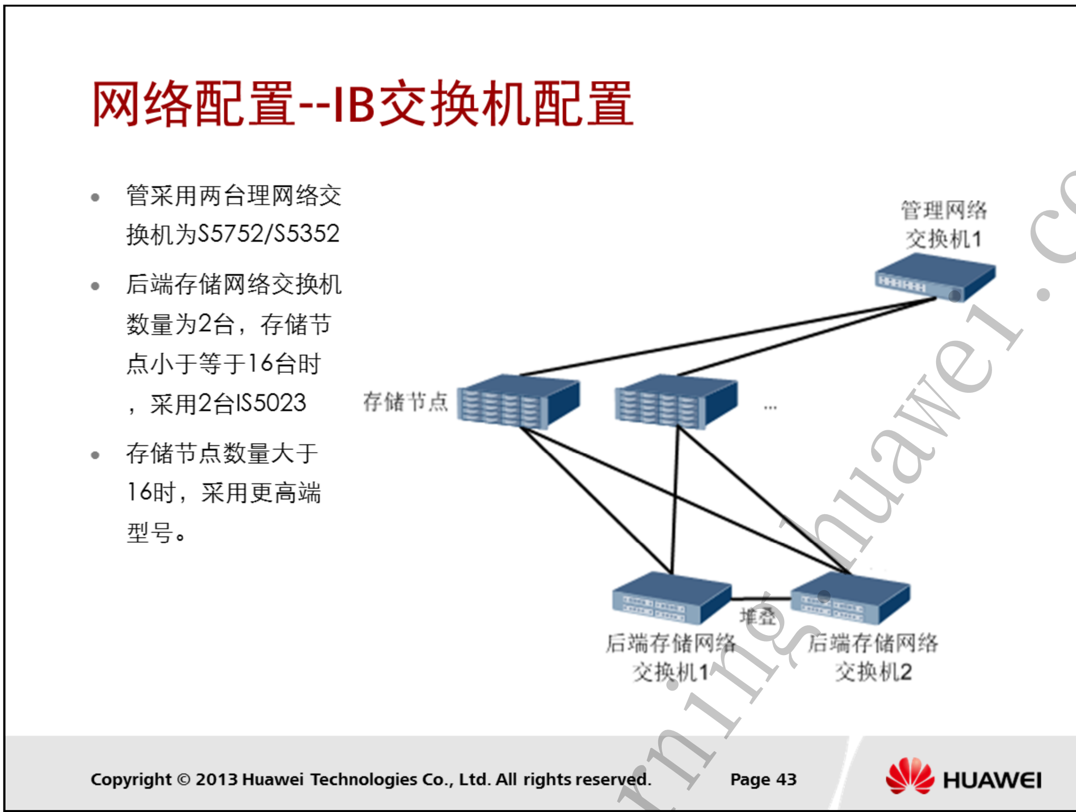
  1、配置两台后端存储网络交换机为堆叠状态。

  2、打开两台后端存储网络交换机中10GE接口的流量控制开关。

  3、为后端存储网络交换机1（高堆叠优先级）的管理网口MEth10/0/0配置一个管理网段的IP地址，然后配置SNMP。

  4、为每台管理网络交换机的VLAN IF1接口配置一个管理网段的IP地址，然后配置SNMP。

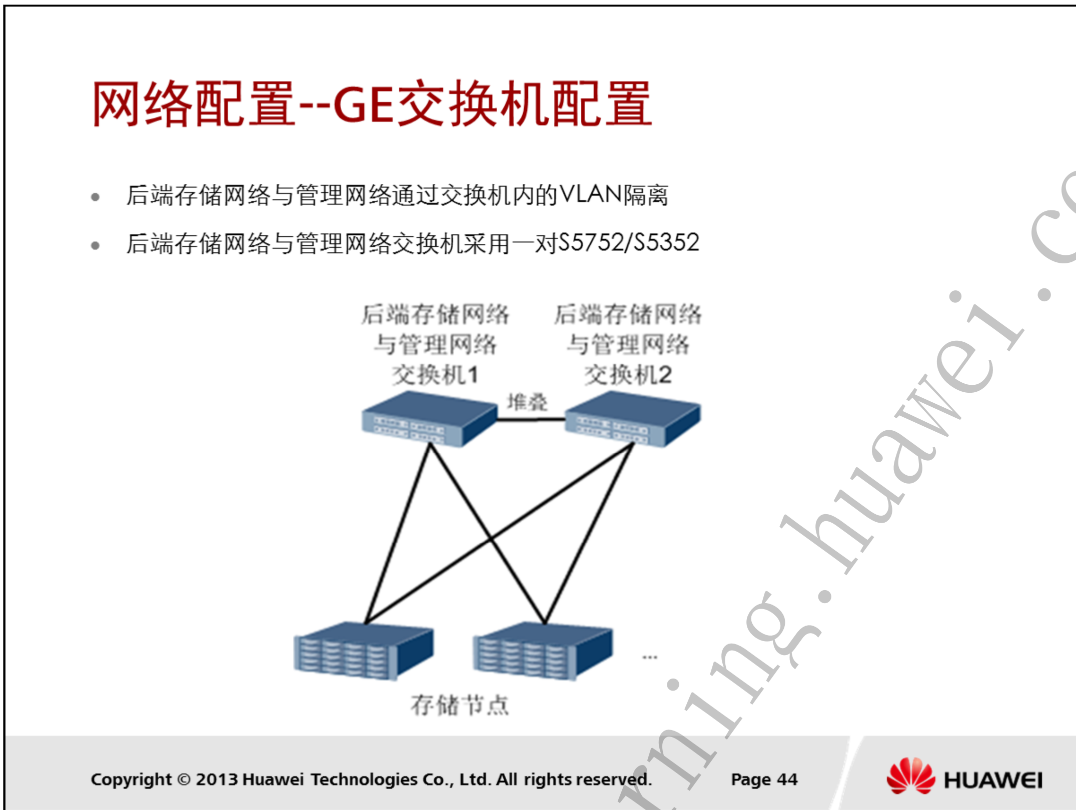
具体操作参见实验手册或产品手册。



- 配置流程：

  1、为每台管理网络交换机的VLANIF1配置1个管理网断IP地址；

  2、为每台管理网络交换机配置SNMP。后端存储网络交换机无需任何配置。



- 配置流程：

  1、配置两台交换机为堆叠状态。

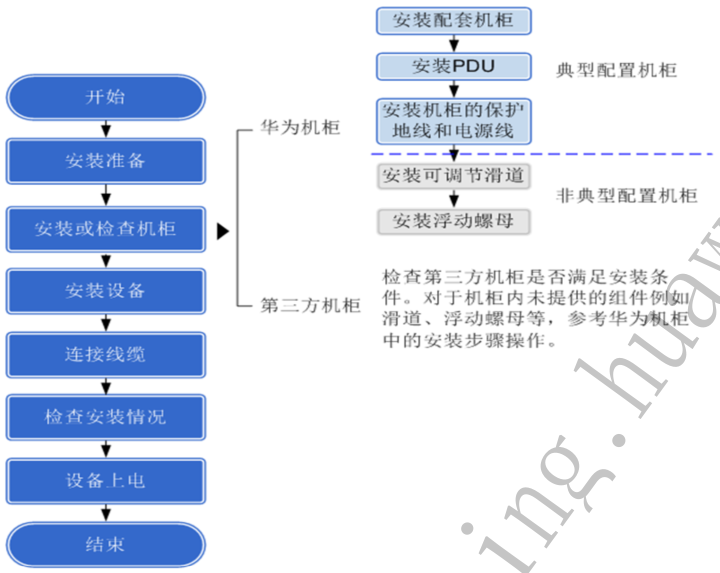
  2、配置VLAN并未管理网络的VLAN的VLANIF接口配置一个管理网络IP地址。

  3、配置SNMP。

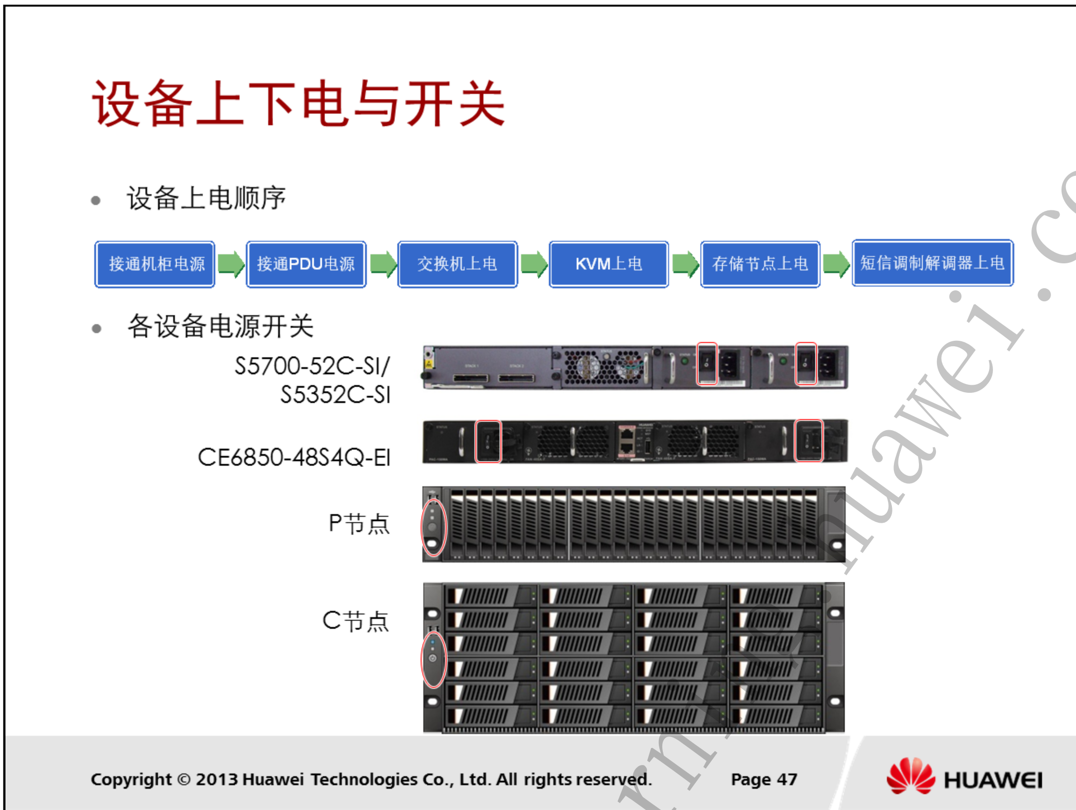
目录

HUAWEI

9000硬件安装流

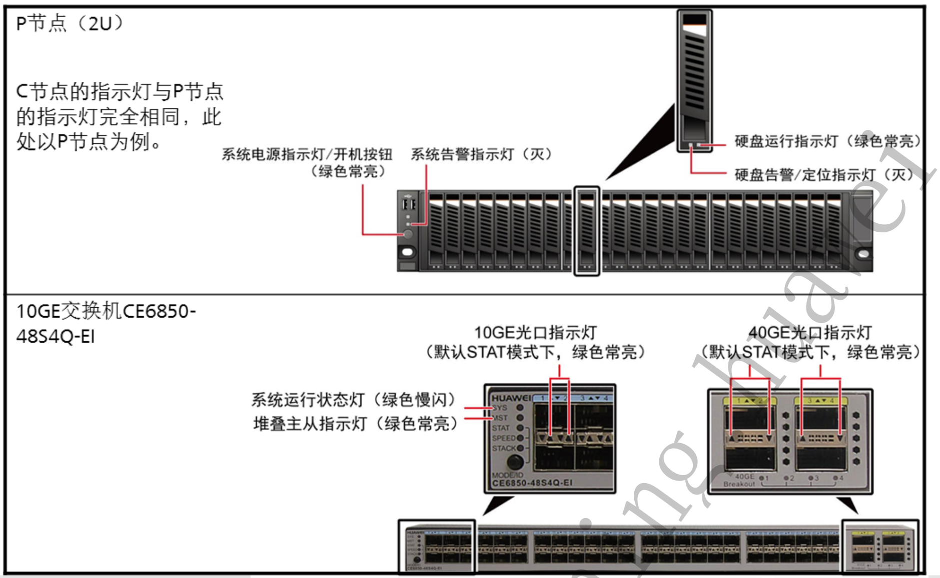


HUAWEI



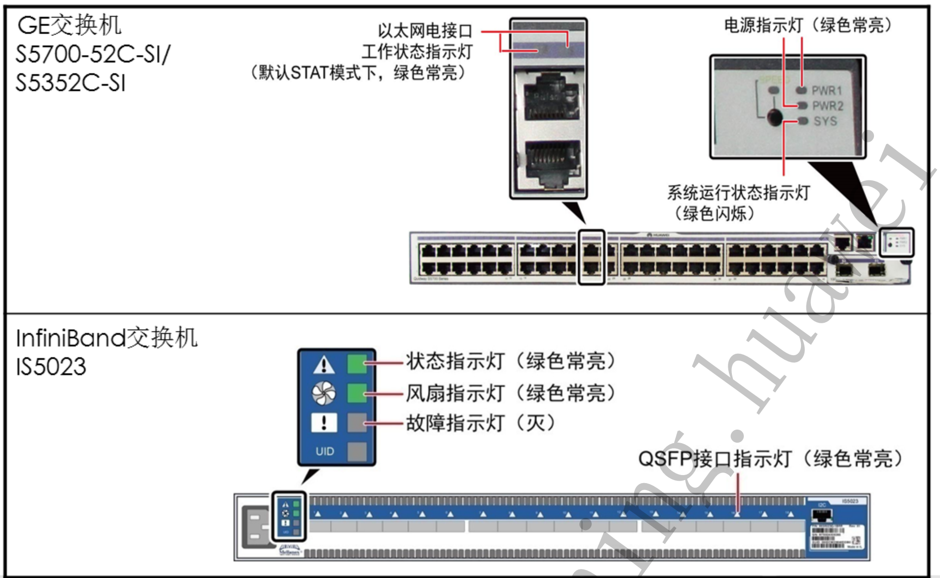
设备下电顺序与上点相反，需要特别注意，在线系统下电，需要先停止上层业务，具体要结合实际情况和方案指导。

系统上电后状态1

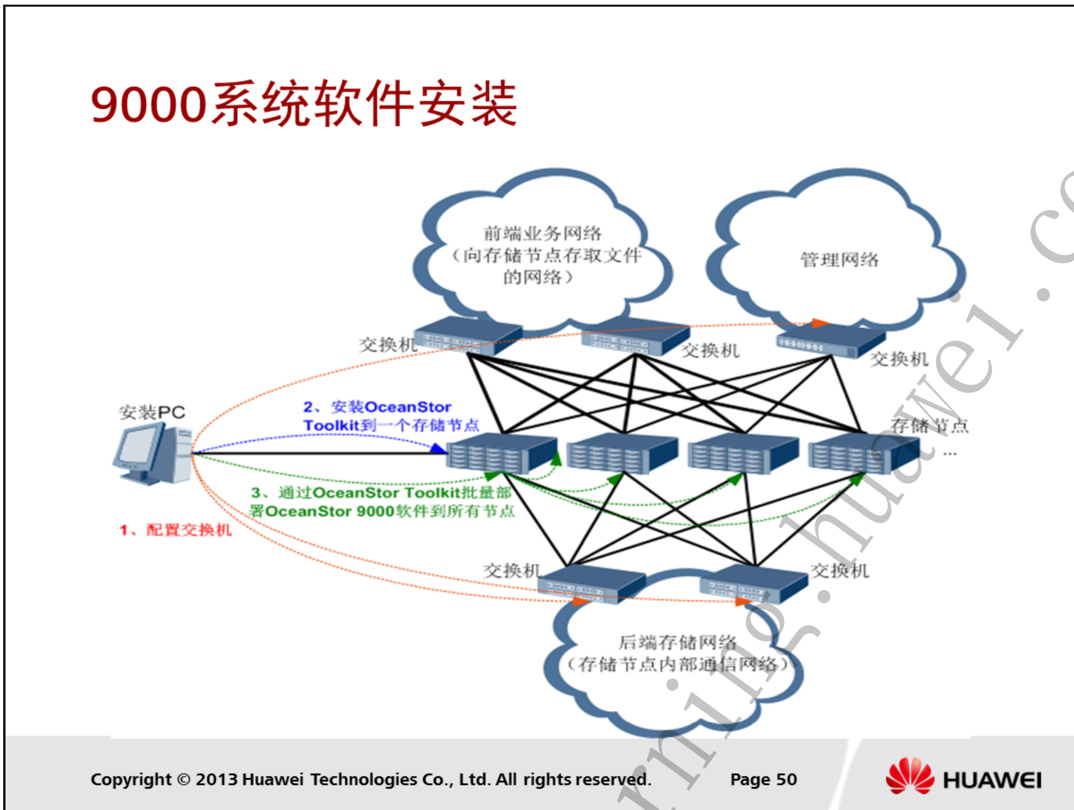


P节点（2U）
C节点的指示灯与P节点的指示灯完全相同，此处以P节点为例。
系统电源指示灯/开机按钮（绿色常亮）
系统告警指示灯（灭）
硬盘运行指示灯（绿色常亮）
硬盘告警/定位指示灯（灭）
10GE交换机CE6850-48S4Q-EI
10GE光口指示灯（默认STAT模式下，绿色常亮）
40GE光口指示灯（默认STAT模式下，绿色常亮）
系统运行状态灯（绿色慢闪）
堆叠主从指示灯（绿色常亮）

系统上电后状态1

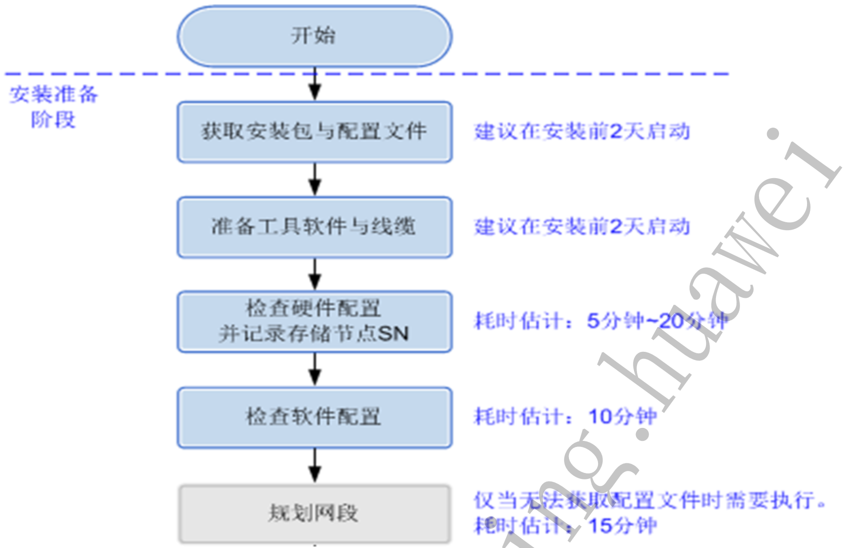


 Page 49 HUAWEI



- OceanStor 9000的软件安装方案要点：
  - 执行软件安装前，应已完成硬件安装与连线。
  - 执行软件安装时，需先配置后端存储网络交换机与管理网络交换机，然后安装OceanStor Toolkit到任意一个存储节点（为方便管理，请选取基础柜中最下方的存储节点），最后应用OceanStor Toolkit通过后端存储网络批量部署OceanStor 9000软件到所有节点
  - OceanStor 9000提供专用性能与容量报表分析工具OceanStor SystemReporter。该报表分析工具为可选安装组件，且无License限制。当管理网络中存在一台满足软硬件配置要求的Windows服务器时即可安装

软件安装流程1



 Page 51 HUAWEI

安装流程2



Page 52 HUAWEI

软件与工具准备

准备项目		说明
软件包	OceanStor ISM-N V100R00xC0x.tar.gz	OceanStor 9000部署工具
	APP_OceanStor 9000_V100R00xC0x.tar.gz	OceanStor 9000系统软件
	mysql-x.x.xx.msi、ISM-N V100R001Cxx.zip	可选报表工具
工具	FTP文件上传工具，如FileZilla	用于上传文件到存储节点。如采用FileZilla，仅需客户端版本，无需Server版本。
	SSH Client软件，如PuTTY	远程访问工具
线缆	一根网线	连接PC到存储设备管理网口。
	一根RS-232串口线缆	连接PC到存储设备串口

 Page 53 HUAWEI

- 软件获取方式：
  - 方式1：随设备发货的软件电子光盘。
  - 方式2：使用企业用户账号和密码登录http://support.huawei.com/enterprise，选择“软件下载 > IT > 存储 > 大数据和云存储 > 大数据存储 > OceanStor 9000”下载左侧所有安装包

获取记录系统组网、节点IP地址等信息的配置文件XXXX.xls。此配置文件应与硬件安装图、线缆连接表等一起，在硬件安装前通过OceanStor 9000网络规划工具统一输出。

软件安装前检查1

置项		要求	检查方法	不符合要求时的处理措施
存储节点[a]	操作系统	SUSE Linux Enterprise Server 11 (x86_64) SP1	1.通过KVM登录存储节点 2.执行命令cat /etc/SuSE-release	安装符合版本要求的操作系统，或联系技术服务支持工程师
	预装部署代理程序	参见《OceanStor 9000版本说明书》	1.通过KVM登录存储节点 2.执行命令rpm -qa | grep deploy	执行升级存储节点中预装的部署代理程序
安装PC	操作系统	Windows 7及以上版本，且已关闭防火墙	-	-
	Web浏览器	Internet Explorer 8及以上版本，或Firefox 11及以上版本	-	-

a：首次安装时，您可以只抽检1～3个存储节点

 Page 54 HUAWEI

- 9000系统存储节点节点数目≥3。如果同时包含P节点（高性能存储节点，2U高度）和C节点（高带宽存储节点，4U高度），则要求两类节点的数目均大于或等于3。
- 在检查时记录每个存储节点前面板铭牌上的设备序列号（显示为“S/N”右侧的20位字符串）到配置文件中。在后续安装过程中，安装程序将以此序列号唯一标识一个存储节点并分配指定的IP地址。
- 如未获取当前项目的配置文件，请记录机柜编号、存储节点编号和设备序列号。
- 当由于高度等原因无法直接查看铭牌时，您可以通过KVM登录存储节点后，执行命令ipmitool fru获取“Product Serial”值。
- 对于已配置IPMI接口IP地址的存储节点，您还可以通过IPMI接口远程登录存储节点GUI管理界面后，在首页查看“设备序列号”值。

安装前检查—控制台要求2

置项		要求	检查方法	不符合要求时的处理
（可选）OceanStor SystemReporter报表服务器	操作系统	Windows Server 2008 R2 Enterprise SP1 64位	-	-
	Web浏览器	Internet Explorer 8及以上版本，或Firefox 11及以上版本	-	-
	TCP端口	下列TCP端口未被占用： 443 80 3306 10091 9616 9002 9003	在Windows命令行窗口中执行命令netstat -ano查看当前端口占用情况。	停止或卸载对应服务。如果443或80端口已被占用，且不便通过上述方法规避，则可在软件安装过程中手工更改为使用其他端口。

HUAWEI



步骤1：用KVM线缆连接KVM至任意一个存储节点（如位于基础柜最下方的节点）。

步骤2：执行命令 iofconfig –a，查看eth0网口的MAC地址。

步骤3：运行yast命令，选择“Network Devices > Network Settings”，敲回车进入。

步骤4：选择一个网口后，查看该网口的MAC地址是否与步骤 2中记录的eth0网口的MAC地址一致，如果不一致，请重新选择另一个网口，直至MAC地址匹配。

步骤5：为此网卡配置一个后端网络网段内的IP地址和子网掩码。

部署工具安装

- 上传后安装成功后登录



  HUAWEI



步骤1：登录部署工具后，单击“部署”。

步骤2：单击上方的“浏览”，选择APP_OceanStor 9000_V100R00xC0x.tar.gz 文件后，单击“导入”。

步骤3：单击“下载模板”，将模板文件下载到本地。

步骤4：根据组网规划结果，编辑模板文件。

数据冗余配比：系统全局缺省冗余配比（部署后管理员可以手工调整任意指定空目录的配比值 ）。缺省为+2，即在将文件切分为数据分条时缺省生成2个冗余分条。SN：贴在设备前面板的铭牌上。

步骤5：单击“导入配置文件”区域中的“浏览”，选择编辑好的文件后，单击“导入”。



步骤1：在“DHCP配置”区域中的网卡中选择部署IP地址，在“IP地址范围”中输入与部署IP地址（说明：部署IP地址是指实际部署时的后端IP地址

）同网段的地址段。注：此IP地址范围仅在部署过程中临时使用。待部署成功后，各节点的IP地址修改为配置文件中规划的IP地址。

步骤2：勾选下方的复选框，连续4次单击“下一步”，然后在“确定启动部署？”对话框中单击“确定”。

配置文件举例1—10GE网络

集群					
集群名称	Cluster				
Link DNS	192.168.100.2	前端网络掩码	255.255.255.224	前端网络网关	192.168.100.1
后端网络掩码	255.255.255.0			数据冗余配比	+2
管理网络掩码	255.255.255.240	管理网络网关	10.110.100.1		
NTP Server地址		管理IP	10.110.100.2		
部署节点					
节点SN	210235G6BQZ0B5000011	节点名称	NODE1	IPMI IP	10.110.100.3
机架号	1	槽位号	1		
网口是否bond	No				
bond0的网口列表					
bond1的网口列表					
bond2的网口列表					
后端IP地址1	172.18.0.1	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.2	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.3	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.4	前端IP地址2的网口	B3-1		
是否独立出管理网口	Yes	OM管理网口	NIC0		
部署节点					
节点SN	210235G6BQZ0B5000012	节点名称	NODE2	IPMI IP	10.110.100.4
机架号	1	槽位号	2		
网口是否bond	No				
bond0的网口列表					
bond1的网口列表					
bond2的网口列表					
后端IP地址1	172.18.0.3	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.4	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.5	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.6	前端IP地址2的网口	B3-1		
是否独立出管理网口	Yes	OM管理网口	NIC0		

 Page 60 HUAWEI

- 集群名称：集群的唯一名称，用于识别集群。由英文字母、数字、“-”和“_”组成，长度在1～15以内的字符串。必须以英文字母开头，不能以“-”或“_”结束。
- InfoLink DNS：OceanStor 9000内用于实现InfoLink特性的DNS代理服务虚拟IP地址。此IP地址应属于前端业务网段。不论是否购买InfoLink特性License，软件安装时均需要填写此IP地址。
- 前端网络掩码、后端网络掩码、管理网络掩码：规划网段 时输出的前端业务网段掩码、后端存储网段掩码与管理网段掩码。
- 前端网络网关：安装程序将此IP地址设置为所有存储节点Linux操作系统的默认网关（Default Gateway）。请联系前端业务网络的维护人员获取此IP地址。一般情况下，当前端业务网络交换机（如果处于透明传输模式，则为相邻的三层交换机或路由器）配置为堆叠状态时，前端业务网络网关应为交换机中与存储节点相连接口所属VLAN的VLANIF地址；当前端业务网络交换机未配置为堆叠状态时，前端业务网络网关应为交换机中与存储节点相连接口所属VRRP Group的Virtual IP地址。
- 管理网络网关：安装程序将此IP地址设置为所有存储节点IPMI接口的默认网关（Default Gateway）。存储节点中IPMI接口的所有数据收发由一个独立服务控制，所以该接口不受Linux操作系统的默认网关影响。
- NTP Server地址：选填参数，NTP服务器的IP地址。使用NTP进行时间同步，可确保所有存储节点中的时间与客户网络中其他设备的时间保持一致。您可以将此参数值保留为空，待软件安装完成后在OceanStor DeviceManager中的“设置 > 集群设置 > 基本信息 > 时间配置”页面中设置NTP Server地址。

配置文件举例2—10GE网络

部署节点					
节点SN	210235G6BQZ0B5000013	节点名称	NODE3	IPMI IP	10.110.100.5
机架号	1	槽位号	3		
网口是否bond	No				
bond0的网口列表					
bond1的网口列表					
bond2的网口列表					
后端IP地址1	172.18.0.5	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.6	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.7	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.8	前端IP地址2的网口	B3-1		
是否独立出管理网口	Yes	OM管理网口	NIC0		
部署节点					
节点SN	210235G6BQZ0B5000014	节点名称	NODE4	IPMI IP	10.110.100.6
机架号	1	槽位号	4		
网口是否bond	No				
bond0的网口列表					
bond1的网口列表					
bond2的网口列表					
后端IP地址1	172.18.0.7	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.8	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.9	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.10	前端IP地址2的网口	B3-1		
是否独立出管理网口	No	OM管理网口			
部署交换机					
IP地址	10.110.100.1				

 Page 61 HUAWEI

- 数据冗余配比：数据切片存储时，冗余条带的个数。此配比值将作为OceanStor 9000中所有目录的默认配比值。对于OceanStor 9000中的空目录，系统允许在部署后随时调整配比值。
  - 可选值包括：
    - +1：适用于对可靠性要求较低的场景。指一个数据分条（Stripe）中有N+1个条带（Strip），可允许任意1个存储节点故障而不影响数据完整性。
    - +2：默认值，适用于大多数场景。指一个数据分条中有N+2个条带，可允许任意2个存储节点故障而不影响数据完整性。此模式所需最少存储节点数量为6，否则将降低冗余度。
    - +3：适用于对可靠性要求更高的场景。指一个数据分条中有N+3个条带，可允许任意3个存储节点故障而不影响数据完整性。此模式所需最少存储节点数量为7，否则将降低冗余度。
    - +4：适用于对可靠性要求最高的场景。指一个数据分条中有N+4个条带，可允许任意4个存储节点故障而不影响数据完整性。此模式所需最少存储节点数量为10，否则将降低冗余度。
    - +2:1：仅当存储节点数目低于20个时考虑。指一个数据分条中有N+2个条带，可允许任意1个存储节点故障而不影响数据完整性。
    - +3:1：仅当存储节点数目低于20个时考虑。指一个数据分条中有N+3个条带，可允许任意1个存储节点故障而不影响数据完整性。

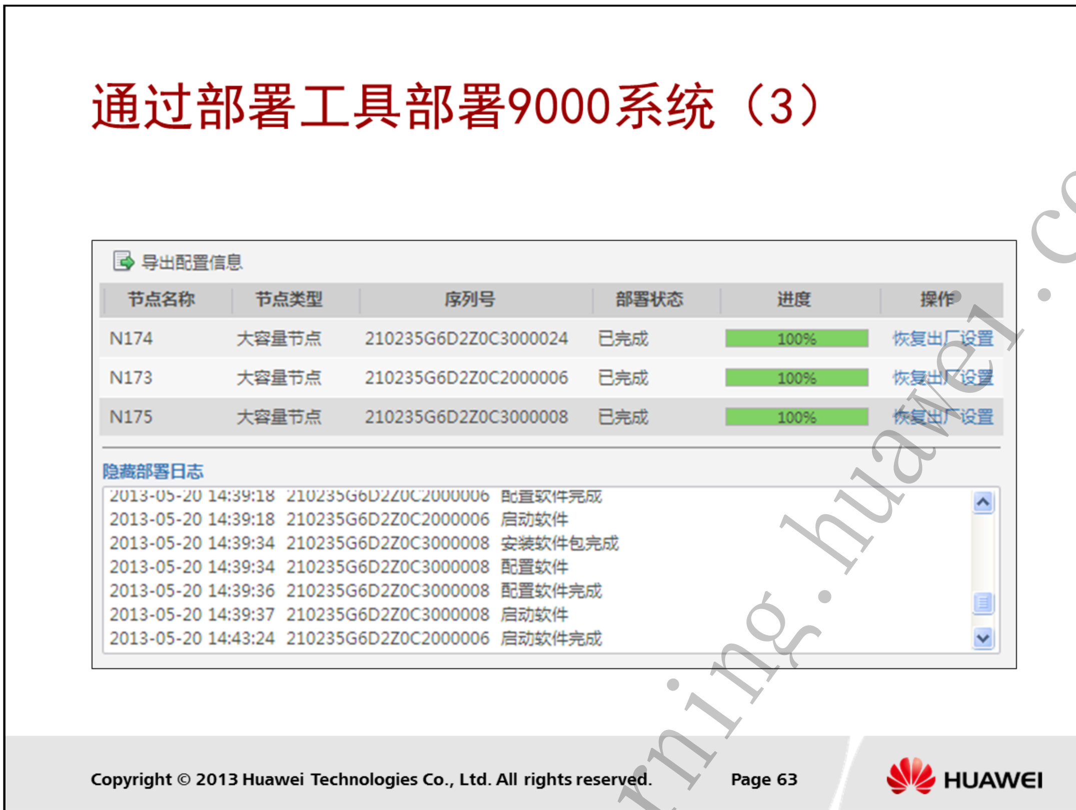
配置文件举例2—10GE网络

部署节点					
节点SN	210235G6BQZ0B5000013	节点名称	NODE3	IPMI IP	10.110.100.5
机架号	1	槽位号	3		
网口是否bond	No				
bond0的网口列表					
bond1的网口列表					
bond2的网口列表					
后端IP地址1	172.18.0.5	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.6	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.7	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.8	前端IP地址2的网口	B3-1		
是否独立出管理网口	Yes	OM管理网口	NIC0		
部署节点					
节点SN	210235G6BQZ0B5000014	节点名称	NODE4	IPMI IP	10.110.100.6
机架号	1	槽位号	4		
网口是否bond	No				
bond0的网口列表					
bond1的网口列表					
bond2的网口列表					
后端IP地址1	172.18.0.7	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.8	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.9	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.10	前端IP地址2的网口	B3-1		
是否独立出管理网口	No	OM管理网口			
部署交换机					
IP地址	10.110.100.1				

 Page 62 HUAWEI

- 上面的N数值为一个小于或等于16的正整数，由系统根据当前存储节点数目自动计算得出。

- 管理IP：OceanStor 9000的管理IP地址，即部署完成后登录OceanStor DeviceManager时所用的IP地址。在10G组网和IB组网中，管理IP属于管理网段；在GE组网中，管理IP属于前端业务网段。



步骤1：系统显示安装进度条，请等待约20分钟。安装成功后，进度显示为100%,状态为已完成。

步骤2：单击“完成”，在弹出的“确认退出部署”对话框单击“确定”，完成OceanStor 9000系统部署。



- 在PC机浏览器输入“https://OM管理IP地址:8088/webui”，登录9000管理控制台（DeviceManager）（户名为admin，默认密码为Admin@storage）
- 如果部署工具无法发现某个节点，请使用KVM登录该节点后，执行命令sh /opt/huawei/deploy/script/deploy_start.sh尝试重启进程。
- 对于因为部署失败或其他原因需要重新部署的节点，请在重新部署前，执行命令/opt/huawei/deploy/script/clean.sh清理环境。

附：9000系统登录信息

系统	入口地址	默认账号/密码	备注
OceanStor Toolkit	https://xx.xx.xx.xx:8088/tools	admin/Admin@storage	完成安装OceanStor Toolkit后可登录
OceanStor DeviceManager	https://xx.xx.xx.xx:8088/webui 其中，“xx.xx.xx.xx”表示基础柜中最下方存储节点NIC0接口的IP地址。此IP地址为验证时临时使用。软件安装完成后，请以配置文件中的“管理IP”代替。	admin/Admin@storage	完成使用OceanStor Toolkit批量部署OceanStor 9000软件后可登录
存储节点IPMI	https://xx.xx.xx.xx。其中，“xx.xx.xx.xx”表示存储节点IPMI接口的IP地址。	admin/Admin@storage	-
存储节点Linux操作系统	-	root/Root@storage	通过KVM登录存储节点或通过PuTTY远程登录存储节点

HUAWEI



登录成功后，如果健康状态、运行状态、版本号、集群容量、节点个数数据正常，无紧急、严重级别告警，则表明软件安装成功。

目录

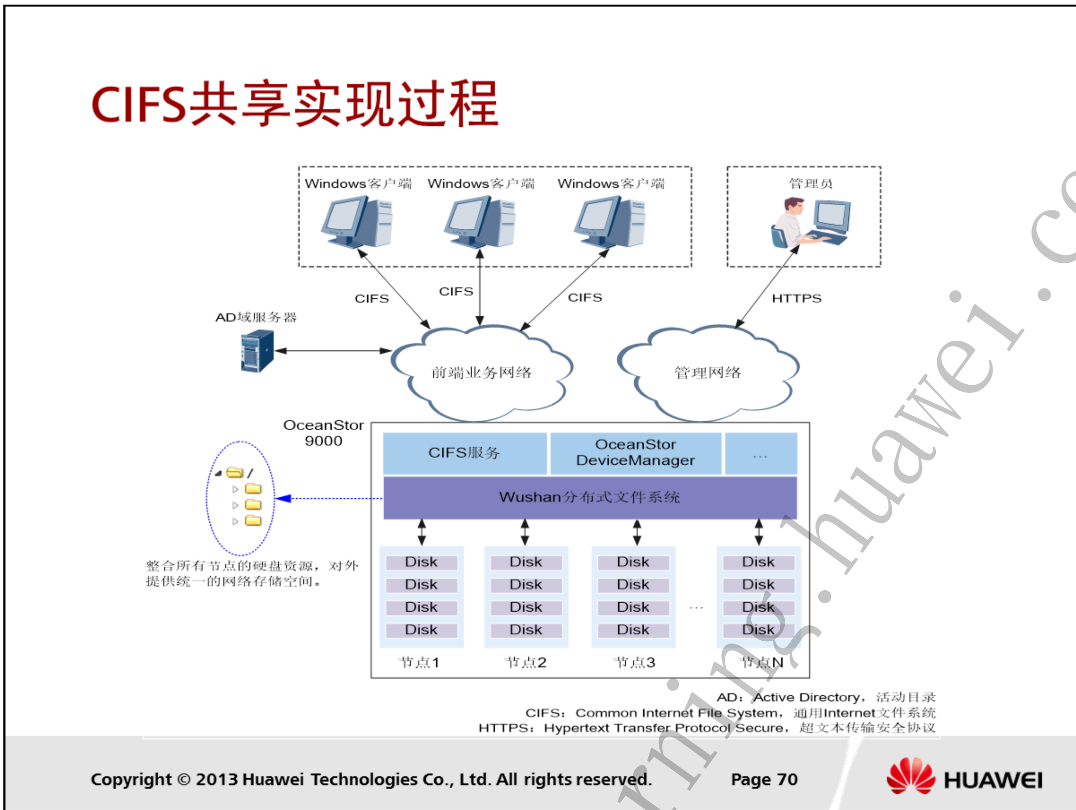
HUAWEI

CIFS/NFS文件共享基本特性

- 同时支持CIFS和NFS协议
- 兼容多种操作系统，可以实现异构网络环境下的文件资源共享
- CIFS/NFS共享的单位是目录，所共享的目录可被多个客户端访问
- 以集群方式对外提供共享服务，节点间可相互监控服务状态
- 可根据业务和节点状态实现负载均衡，数据访问在集群内均匀分布
- NFS支持V3，V4协议
- SMB支持V1.0，V2.0

CIFS支持的认证方式

- 本地用户认证
  - 由OceanStor 9000本地进行接入认证，允许规划的本地用户/用户组接入。当启用了Guest用户接入权限后， 允许用户匿名访问共享目录，即用户无需输入用户名和密码即可接入OceanStor 9000
- AD（Active Directory）域用户认证
  - 用户网络中配置AD域服务器，由AD域服务器进行接入认证，允许合法的AD域用户接入

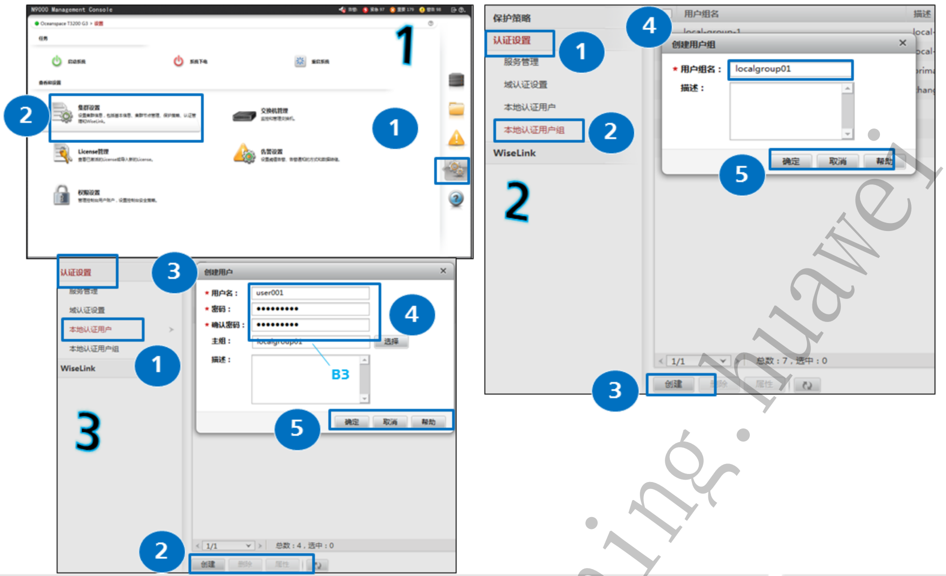


OceanStor 9000运行时，其内置的Wushan分布式文件系统自动整合所有节点内的硬盘资源，并对外显示为一个根目录。要使一个或多个Windows客户端访问OceanStor 9000内的存储资源，管理员仅需要在根目录下创建待访问的目录，并将其授权给目标用户或用户组。

OceanStor 9000同时支持本地认证与AD域服务器认证两种方式。采用本地认证时，管理员应提前在OceanStor DeviceManager中添加本地认证用户组与用户；采用AD域服务器认证时，管理员应提前将OceanStor 9000加入域。

- 配置流程：
  - 登陆OceanStor 9000管理控制台
  - 创建用于共享的目录
  - 创建认证方式(本地用户认证/AD域认证)
  - 创建CIFS共享

添加本地认证用户组与用户（可选）



HUAWEI

加入域（可选）

- 先通过OceanStor DeviceManager登录系统
- 域环境下，需要先将OceanStor 9000加入域环境



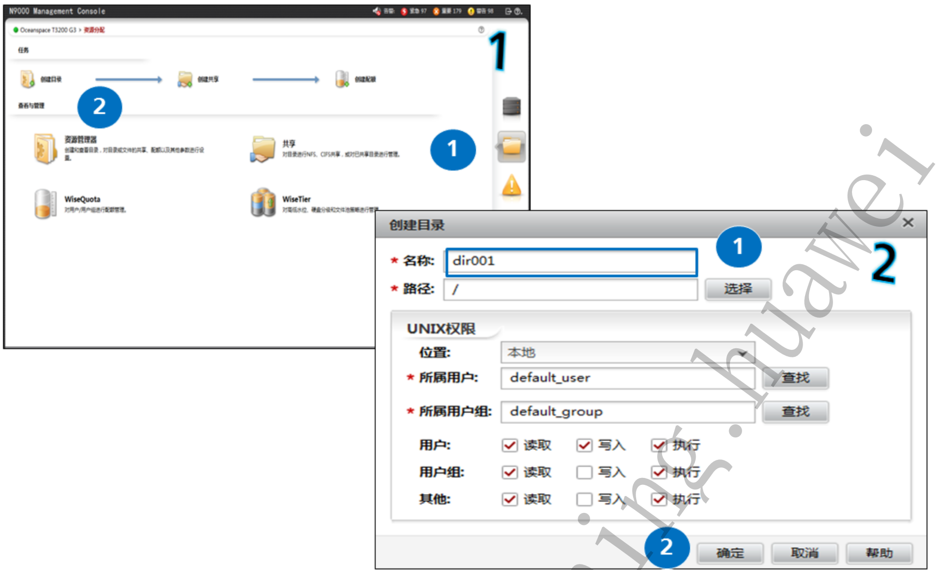
 Page 72 HUAWEI

如果提示加入失败，请确认已在“设置 > 集群设置 > 基本信息 > 外部DNS配置”页面中配置外部DNS服务器的IP地址，且确保网络互通。

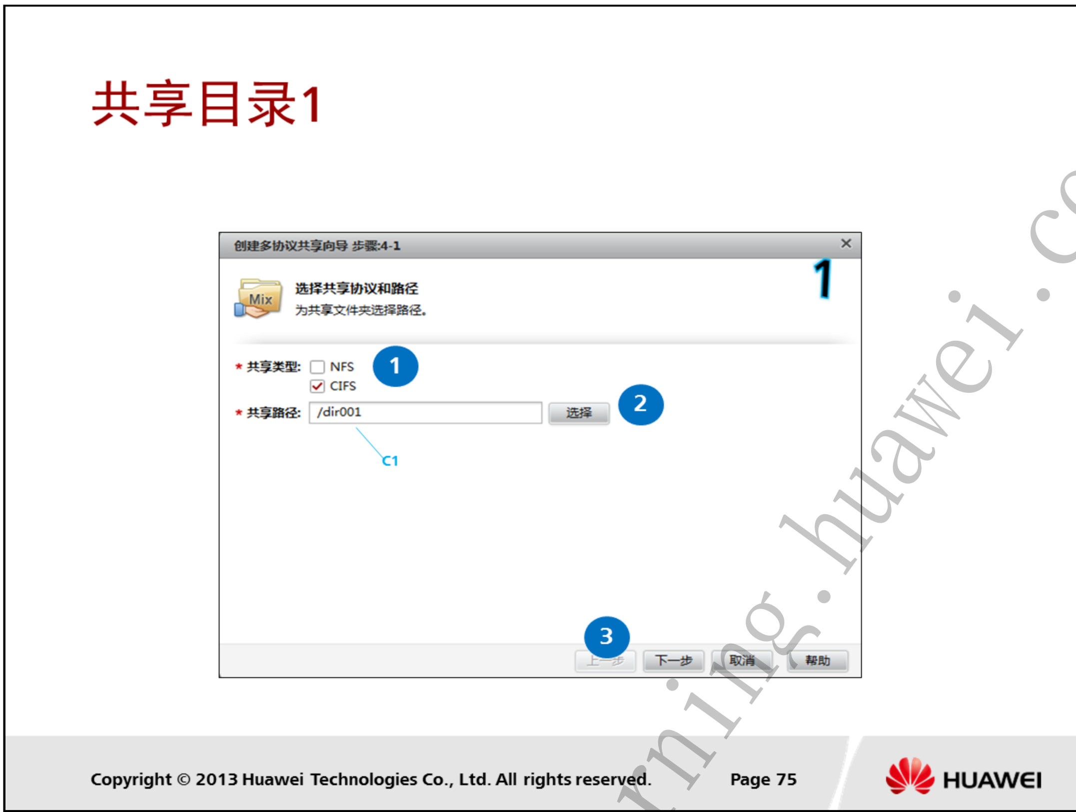


- 在无域环境中，请设置“认证模式”为“用户模式”； 在AD域环境中，请设置“认证模式”为“域模式”。

创建目录

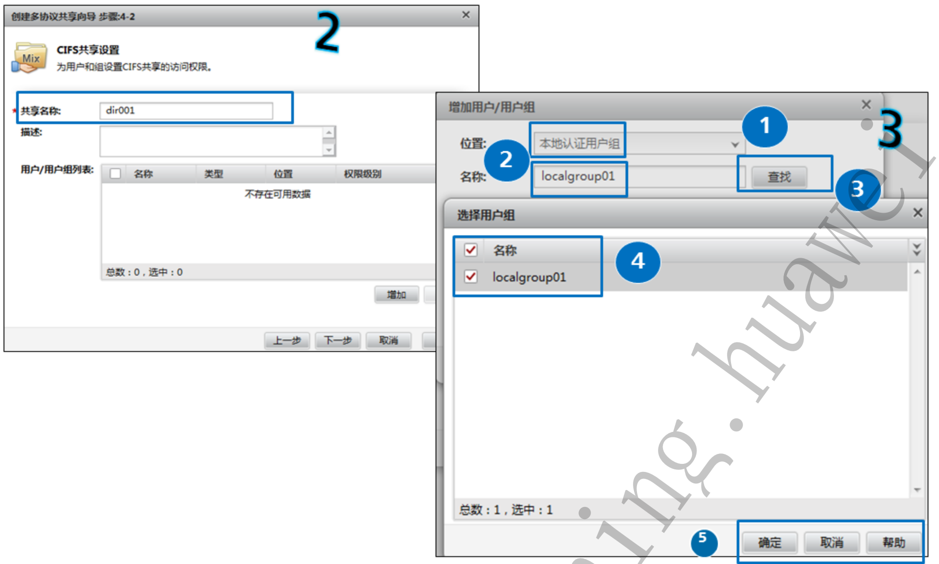


创建目录
名称:
dir001
路径:
/
选择
UNIX权限
位置:
本地
所属用户:
default_user
查找
所属用户组:
default_group
查找
用户:
读取
写入
执行
用户组:
读取
写入
执行
其他:
读取
写入
执行
确定
取消
帮助

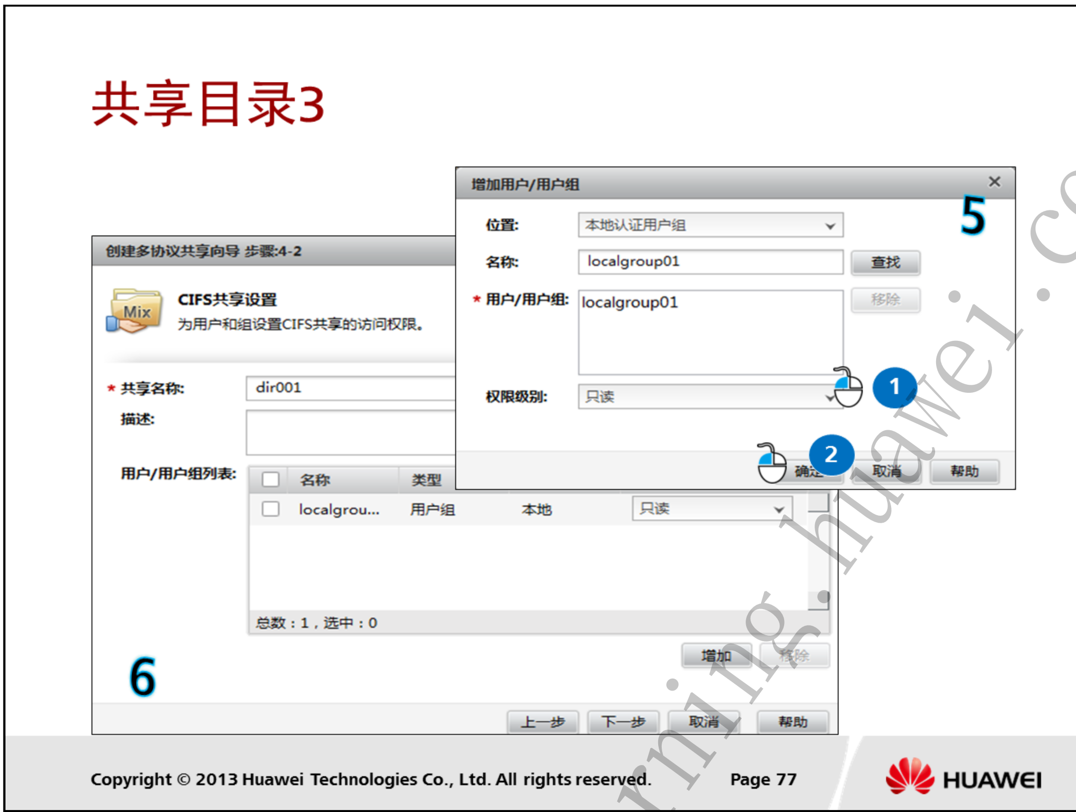


- 在AD域环境中，如果不启用本地用户认证，则无需执行 3、4 、5 步骤

共享目录2



 Page 76 HUAWEI


共享目录3
创建多协议共享向导 步骤:4-2
Mix
CIFS共享设置
为用户和组设置CIFS共享的访问权限。
* 共享名称:
dir001
描述:
用户/用户组列表:
名称
类型
localgrou...
用户组
本地
只读
总数：1，选中：0
增加
移除
6
上一步
下一步
取消
帮助
增加用户/用户组
位置:
本地认证用户组
名称:
localgroup01
查找
* 用户/用户组:
localgroup01
移除
权限级别:
只读
5
1
2
确定
取消
帮助
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 77
HUAWEI

共享目录4



 Page 78 HUAWEI

共享目录5



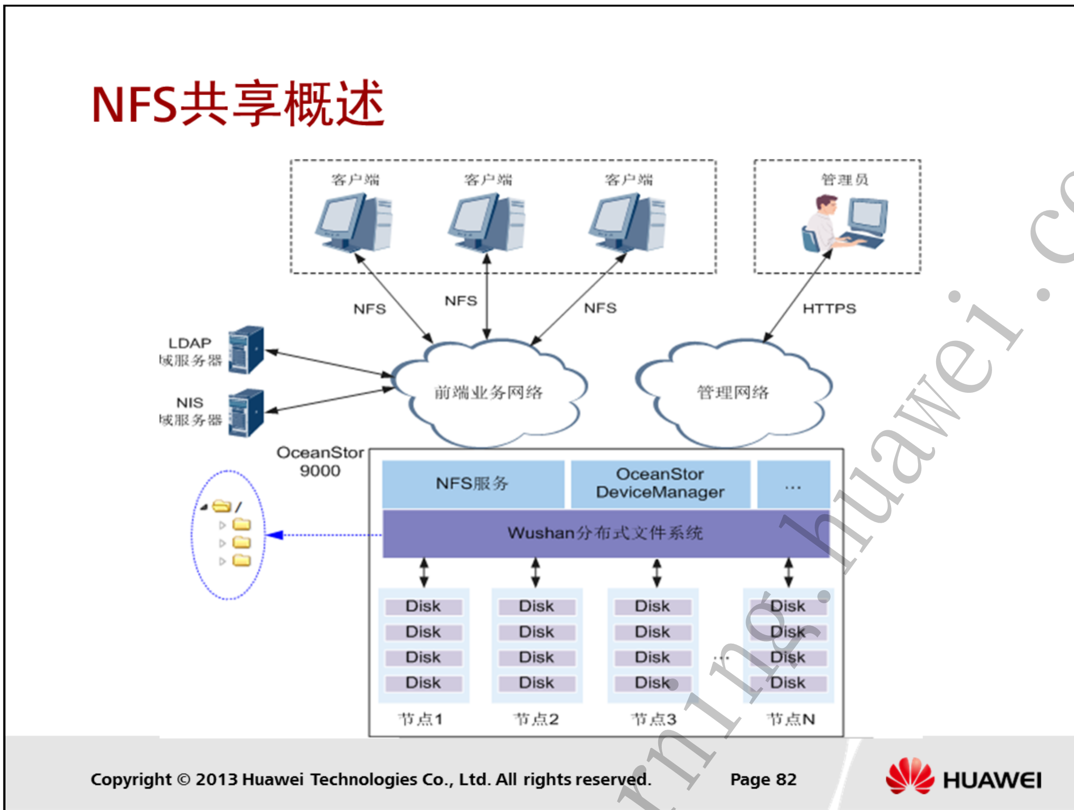
HUAWEI



- 请确保当前Windows客户端与OceanStor 9000前端业务网络之间网络互通。
- 下面以Windows 7操作系统为例介绍访问方法。

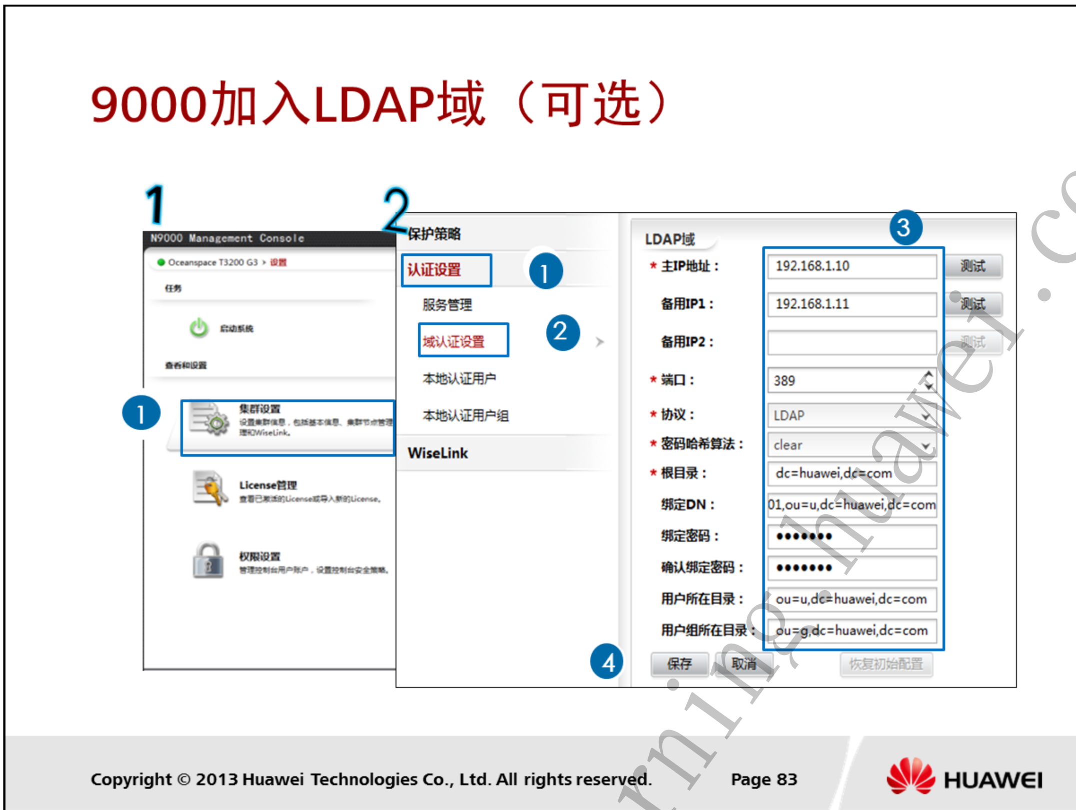


- 全部操作已完成，就可以像使用本地文件系统一样使用OceanStor 9000提供的存储空间。

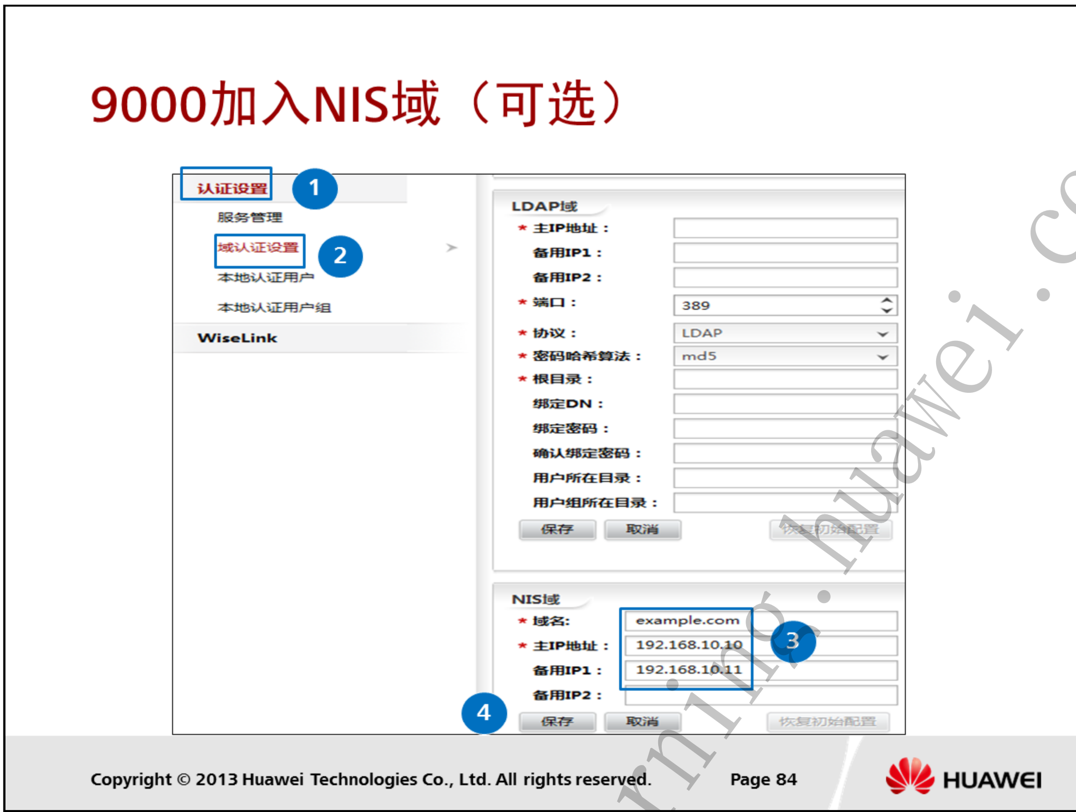


OceanStor9000运行时，其内置的Wushan分布式文件系统自动整合所有节点内的硬盘资源，并对外显示为一个根目录。要使一个或多个Windows客户端访问OceanStor9000内的存储资源，管理员仅需要在根目录下创建待访问的目录，并将其授权给目标用户或用户组。

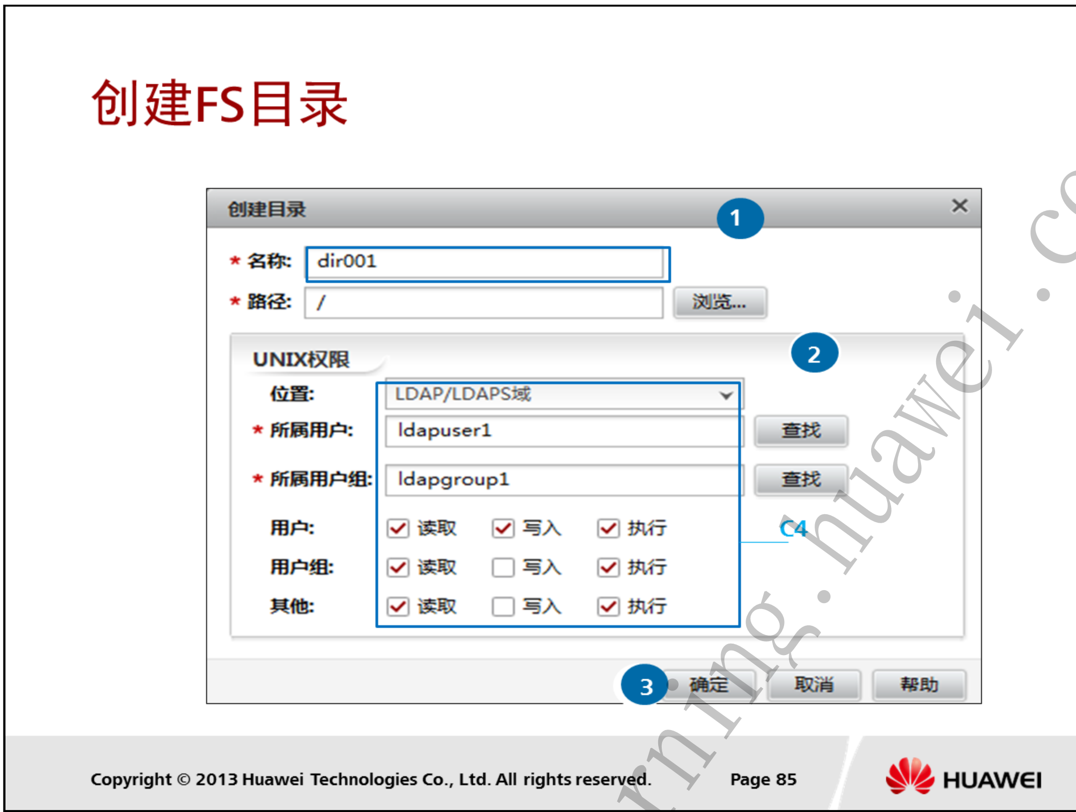
- 配置流程：
  - 建立9000与客户端网络，基于TCP/IP的NFS通信正常
  - 登陆OceanStor 9000管理控制台，配置文件系统
  - 创建用于共享的目录
  - 配置认证方式(LDAP认证/NIS认证)
  - 在客户端访问NFS共享



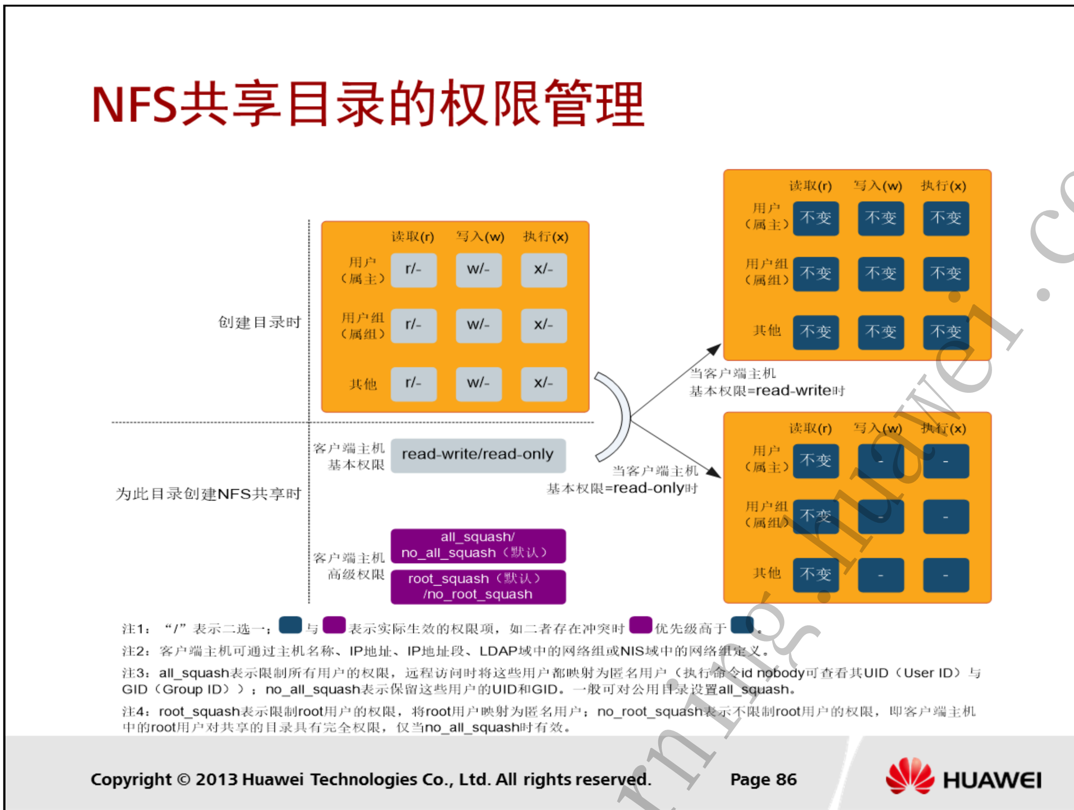
- 在“LDAP域”区域输入主用LDAP服务器的IP地址。
- （可选）分别输入备用LDAP服务器1和2的IP地址。如果存在备用LDAP服务器，则需要执行此步骤。
- 输入OceanStor 9000与LDAP服务器通信所使用的端口号。端口号与LDAP服务器侧配置为一致。默认端口号为389。
- 选择OceanStor 9000与LDAP服务器通信使用的协议。如果LDAP服务器支持SSL，则可以选择LDAPS。
- 选择报文的加密算法。OceanStor 9000支持MD5、Crypt和Clear加密算法。
- 输入LDAP目录的基准DN。首字母不能是空格和“#”，尾字母不能是空格，格式为xxx=yyy，使用逗号分隔。
- 输入绑定DN。首字母不能是空格和“#”，尾字母不能是空格，格式为xxx=yyy，使用逗号分隔。
- 输入密码。登录LDAP服务器的密码，需要输入两次进行确认，两次密码输入要求一致。
- 输入一个用户DN。用户目录项。
- 输入一个用户组DN。用户组目录项。
- 单击“保存”，完成加入LDAP域设置。



- 在“NIS域”区域输入域名。例如：nisserver.com
- 设置NIS域的域名。输入主用NIS服务器地址。例如：192.168.10.10
- 设置NIS服务器的IP地址。（可选）分别输入备用NIS服务器1和2的IP地址。如果存在备用NIS服务器，则需要执行此步骤。
- 单击“保存”。完成加入NIS域设置。

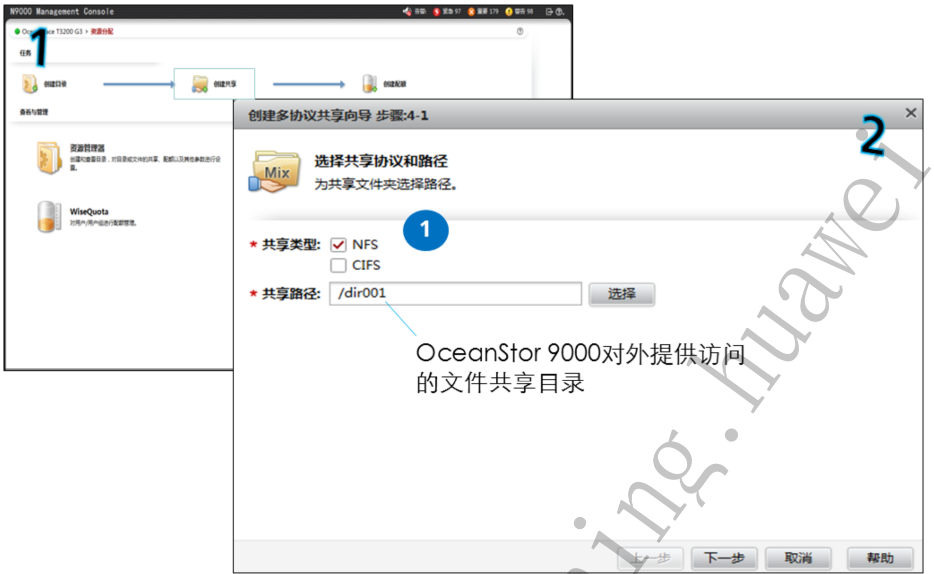


- 配置入口：OceanStor 9000管理控制台>资源分配（界面右侧）>资源管理器
- C4 设置共享目录的所属用户、用户组及权限

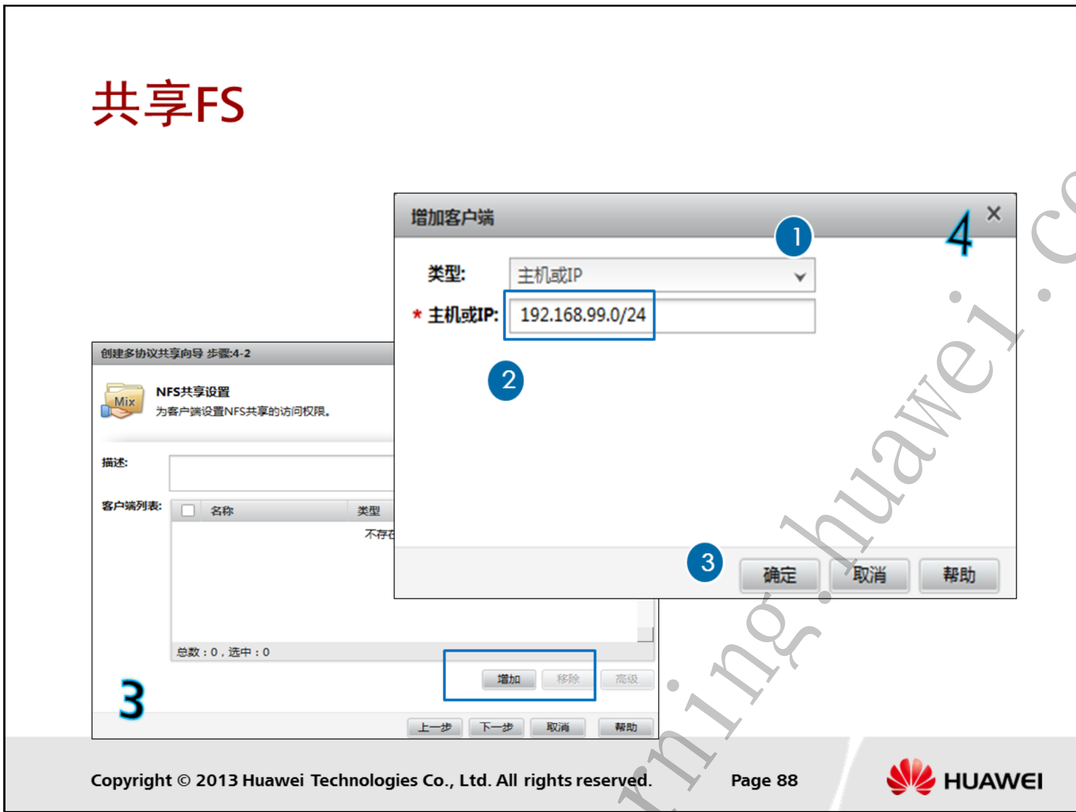


- 在LDAP域或NIS域环境中，属主、属组可指定为域用户、域用户组，因此您可以在客户端主机中以此域用户或此域用户组身份登录并以属主或属组权限访问共享目录。
- 在无域环境中，客户端主机中的所有用户均以其他用户权限访问共享目录。

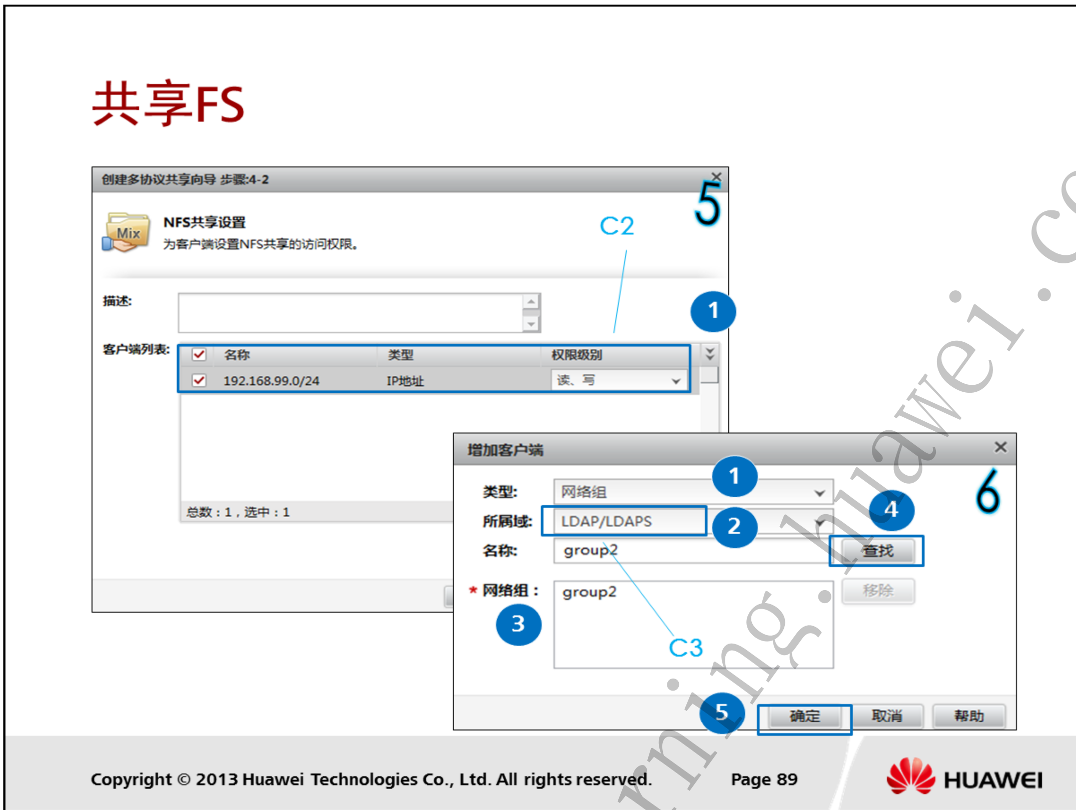
共享FS



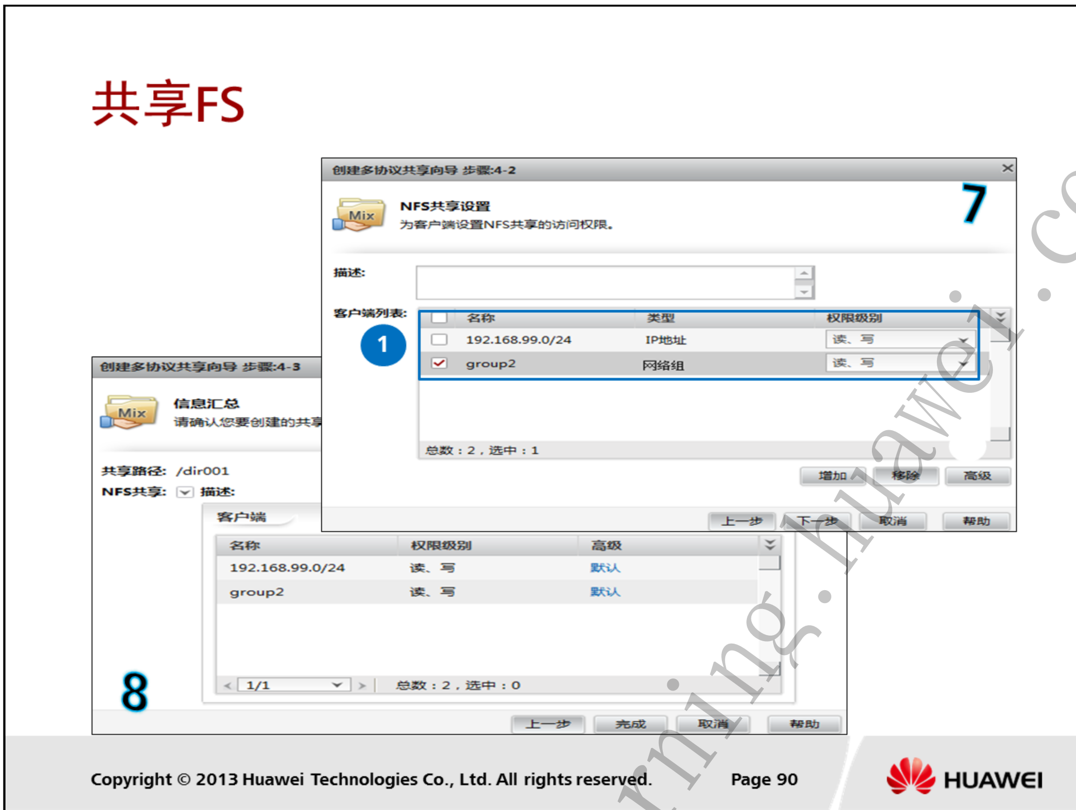
创建多协议共享向导 步骤:4-1
选择共享协议和路径
为共享文件夹选择路径。
共享类型:
NFS
CIFS
共享路径:
/dir001
选择
OceanStor 9000对外提供访问的文件共享目录
上一步
下一步
取消
帮助



- 如果待添加的客户端的类型为主机名称，则选择“主机名”
- 如果待添加的客户端的类型为IP地址，则选择“IP地址”
- 如果网络环境中部署了LDAP或NIS域服务器，待添加客户端的类型为网络组，则选择“网络组”



- 要在客户端中允许以root用户访问共享目录，请单击“高级”，在“root权限控制”中选择“no_root_squash”
- C2 C3 指定访问文件共享目录的客户端范围及权限



- 要在客户端中允许以root用户访问共享目录，请单击“高级”，在“root权限控制”中选择“no_root_squash”



- 前提确保当前客户端与OceanStor 9000前端业务网络之间网络互通

操作系统类型	挂载命令	举例
SUSE、Red Hat	mount -t nfs *ipaddress:sharedir localdir*	mount –t nfs 192.168.100.3:/dir001 /mnt
Mac OS	sudo mount -t nfs *ipaddress:sharedir localdir*	sudo mount -t nfs 192.168.100.3:/dir001 /Volumes/mount1

目录

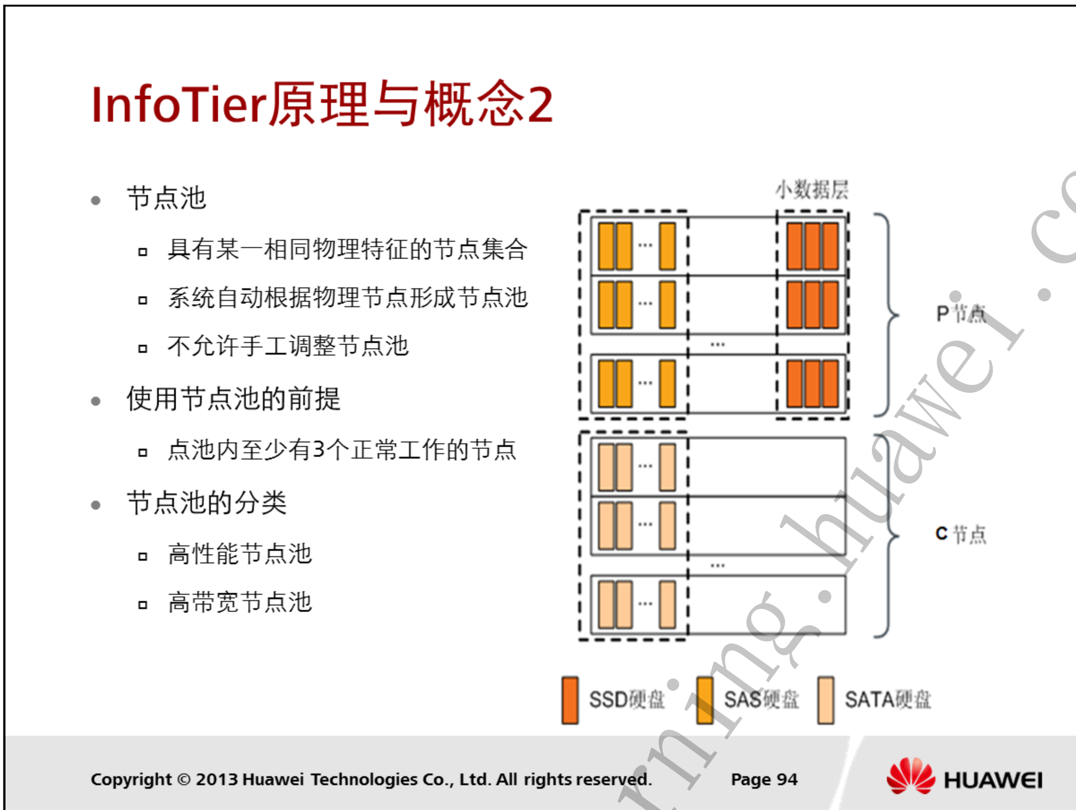
InfoTier特性

- InfoTier特性包括的主要功能如下：
  - 存储级别：在OceanStor 9000内部根据节点类型和硬盘类型对硬盘分级，相同等级的硬盘组成一个存储级别。
  - 高低水位策略：高低水位是指存储级别中空闲空间占用硬盘存储总容量的百分比，OceanStor 9000通过设置全局统一的高低水位，限制每个存储级别中的文件迁移，避免硬盘存储容量不足造成的故障。
  - 文件池策略：通过设置文件属性与各个存储级别间的对应关系，定义文件在不同存储级别中存储和迁移的策略。

 Page 93 HUAWEI

InfoTier特性也称为自动分级存储，可将存储空间划分为不同的存储级别，根据文件属性的区别将文件存放在不同的存储级别上，满足用户对文件处理速度、传输带宽和存储容量的多样需求，以达到合理利用存储空间、提升存储系统访问性能、降低总体部署成本的目的。

本特性是OceanStor 9000的可选特性，需要获得License许可才能获得该特性的服务。当未购买该特性的License时，OceanStor 9000只提供硬盘分级功能，不支持自动迁移文件存储位置，管理员无法设置文件池策略和高低水位。仅当OceanStor 9000内部存在不同类型的节点，并且每种类型的节点数量大于或等于3个时，才可以开启分级存储特性实现自动迁移文件存储位置功能。



- 高性能层：由P节点SAS硬盘组成的存储级别，通常用于存储用户访问频度较高、响应速度要求较快的文件。
- 高带宽层：由C节点SATA硬盘组成的存储级别，通常用户存储用户访问频度较低、传输带宽和存储容量要求较高的文件。
- 小数据层：由P节点SSD硬盘组成的存储级别，通常用于存储小文件，文件处理响应速度最快，但存储容量最小。
- 高性能节点池
  - 由P节点组成，应用于文件读写频繁且对读写速度要求较高的OPS密集型应用场景
- 高带宽节点池
  - 由C节点组成，应用于对文件读写带宽要求较高的高带宽应用场景
- 由于SSD硬盘容量较小，建议存放于小数据层的文件不大于1MB。
- P节点配置4块SSD硬盘，C节点配置1块SSD硬盘。但由于OceanStor 9000软件固定占用一块SSD硬盘，因此C节点的SSD硬盘无法加入到小数据层，只有P节点可以将3块SSD硬盘加入小数据层。

InfoTier原理与概念3

- 文件池
  - 具有共同特征的文件集合
  - 各个特征之间是“与”的关系
  - 同一个文件可以属于多个文件池
- 文件池策略
  - 文件池如何在节点池存放的规则
  - 文件池策略具有优先级
  - 购买license才允许配置
- 文件池策略的使用
  - 系统中最多可配置128条文件池策略
  - 文件对应的最高优先级的文件池策略才会生效

HUAWEI

InfoTier原理与概念4

- 高低水位
  - 节点池中空闲空间占用总空间的百分比
  - 可在ISM上配置，对所有节点池生效
  - 高水位默认为10%，低水位默认为5%
- 高低水位使用场景
  - 处于低水位以下时，禁止文件写入/迁入该节点池
  - 恢复到高水位时，才允许文件迁入该节点池
  - 处于低水位以上时，允许文件写入该节点池

HUAWEI

文件初始存放

- 匹配文件池策略
  - 只匹配文件池特征包含且仅包含文件路径的文件池策略
  - 如果同时匹配到多个文件池策略，选择优先级最高的
  - 通过文件池策略确定存放的节点池； 如未匹配到，选择空闲空间最大的
- 确定节点池
  - 查询对应的节点池当前空闲容量
  - 高于低水位，使用该节点池
  - 低于低水位，使用系统中空闲空间最大的节点池
  - 节点池全部处于低水位以下，拒绝该文件的写入

HUAWEI

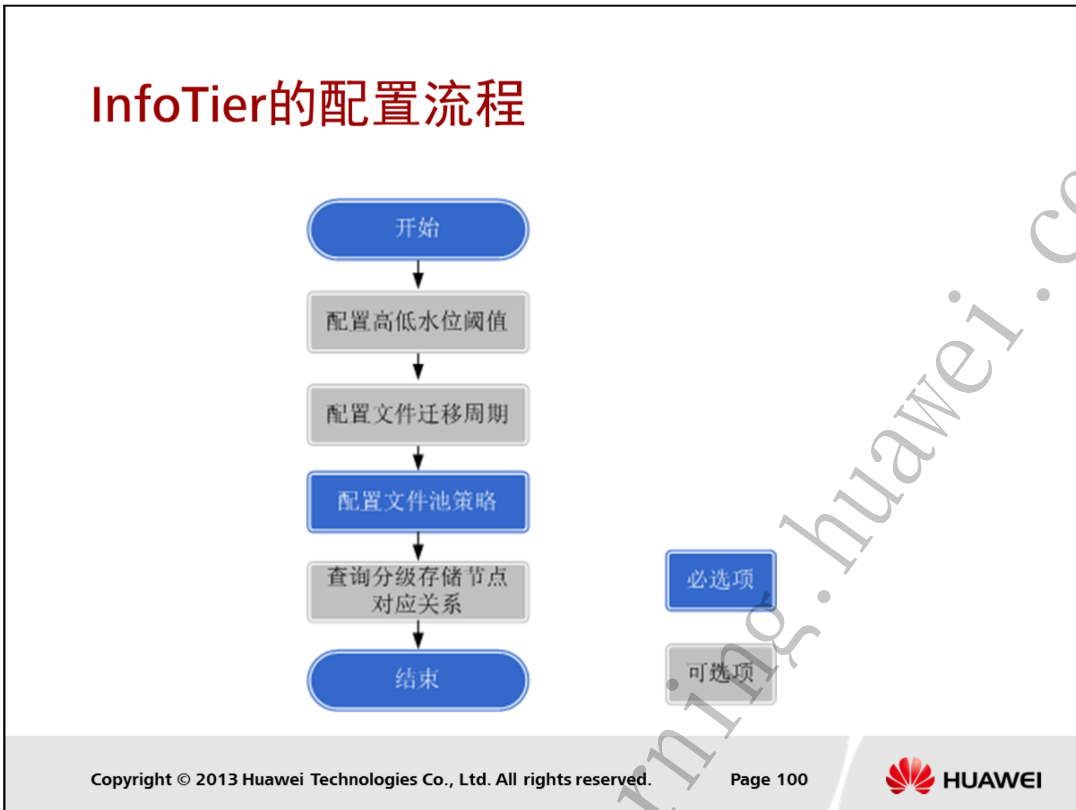
文件迁移

- 迁移扫描
  - 扫描起始和结束时间可配置，缺省关闭扫描
  - 读取系统内文件的属性，得到文件当前所在的节点池
  - 匹配系统内的文件池策略，确定文件的目标节点池
  - 判断文件是否需要迁移，生成迁移任务
- 文件迁移
  - 逐chunk进行迁移
  - 在迁移过程中，不影响文件的正常读写
  - 迁移前后，文件的保护级别不会降低
  - 目标文件池空闲空间高于低水位

InfoTier的配置

- 配置之前的准备工作
  - 确认是否购买了license
  - 是否存在不同类型的节点，且每种类型的节点至少3个
  - 配置规划
  - 文件初始存放位置
    - 有明确规划的按照路径配置文件池策略
    - 无明确规划的不配置对应的文件池策略，由系统自动分配
  - 文件后期存放位置
    - 确定节点池对应文件应具有的属性：热度，存放路径
    - 根据文件的属性配置对应的文件池策略

  HUAWEI



- 配置高低水位阈值:管理员通过合理配置高低水位策略，限制不同存储级别间的文件写入和迁移操作，确保存储空间利用率处于合理的水平。
- 配置文件迁移周期:管理员通过配置文件迁移周期，可使文件属性扫描在业务量较小的时间段进行，避免对存储系统业务性能造成影响。
- 配置文件池策略:管理员通过合理配置文件池策略，充分发挥分级存储特性的能力，激活文件在不同存储级别间的自动迁移，提高系统存储性能，降低总体存储成本。
- 查询分级存储节点对应关系:管理员配置分级存储策略后，可查询不同存储分级与物理节点的对应关系。

高低水位的配置

参数名称	参数说明
高水位	通过高水位阈值定义存储分级空闲容量超过规定数值后，是否可以恢复数据写入或迁入操作。
低水位	通过低水位阈值定义数据是否可以初次写入或迁入指定的存储分级。

HUAWEI

- 系统默认的高水位是空闲空间占用存储分级硬盘总容量的10%，管理员可以修改高水位阈值。
- 当存储分级的空闲空间从低水位恢复到高水位时，将允许文件迁入。
- 系统默认的低水位阈值是空闲空间占用存储分级硬盘总容量的5%，管理员可以修改低水位阈值。 当文件初次写入指定存储分级时，如果该分级的空闲空间小于低水位阈值，则不能写入该分级，而是选择空闲空间最大的分级写入。
- 当文件在迁移到指定存储分级的过程中，如果该分级的空闲空间小于低水位阈值，则禁止文件迁入。

扫描时间配置

参数名称	参数说明
起始时间	扫描节点时，从监控模块读取文件池策略、文件扫描起始时间和结束时间，到达起始时间时系统自动启动扫描
结束时间	扫描节点时，从监控模块读取文件池策略、文件扫描起始时间和结束时间，到达结束时间时无论扫描是否全部完成系统均强制停止扫描操作

HUAWEI

- 起始时间的最小单位可设置到分钟，且起始时间和结束时间最少需间隔1小时，即若起始时间设置为0点，结束时间需设置为1点及其后时刻。
- 起始时间默认设置为0时，即每日0时自动启动文件属性扫描。可根据局点情况设置为系统访问较空闲的时间。
- 结束时间的最小单位可设置到分钟，且起始时间和结束时间最少需间隔1小时，即若起始时间设置为0点，结束时间需设置为1点及其后时刻。
- 结束时间与起始时间相同认为关闭扫描

文件池策略的配置

参数名称	参数说明
名称	通过文件池策略的名称标识不同的文件属性扫描规则。
分级	通过级别定义每个文件池策略对应的存储分级。
件名	通过定义文件名称中具有代表意义的部分，指定具有相同属性文件的存储策略。
文件路径	通过合理规划文件系统目录结构，定义特定路径对应的存储级别。
文件大小	通过文件大小定义对应的存储级别
修改时间	通过文件最近一次修改时间与当前时间的间隔定义对应的存储级别
创建时间	通过文件创建时间与当前时间的间隔定义对应的存储级别
最后一次访问时间	通过文件最近一次访问时间与当前时间的间隔定义对应的存储级别
I/O次数	通过监控在统计周期内对文件的访问次数定义对应的存储级别
I/O热度	通过监控在统计周期内对文件的访问热度定义对应的存储级别

 Page 103 HUAWEI

- I/O次数表示在指定的统计周期内的访问操作次数，可定义的统计周期为1天以内～7天以内
- I/O热度表示在指定天数内的数据访问热度，可定义的统计周期为1天以内～7天以内。
- I/O热度的计算公式为“I/O热度=（访问字节数/文件大小）*100”，表示统计周期内文件访问字节数占文件自身大小的百分比。设置参数时只需输入数值即可，不需要输入“%”

目录

HUAWEI

InfoQuota技术背景

- 配额技术是一种基于用户、用户组文件存储管理
- 还支持目录的配额管理
- 使得管理员可以方便合理地分配存储资源
- 该功能受License控制

InfoQuota原理

- 以用户和用户组统计在特定目录下的文件占用空间
- 在内核文件子系统上实现
- 根据文件读写创建请求分析出本次操作受影响的配额控制项
- 逐项进行配额的门限值比较，根据门限值允许或禁止本次文件操作
- 如果受影响的所有配额控制项允许，请求继续进行，并更新受影响的配额项
- 只要其中一项不允许则返回失败

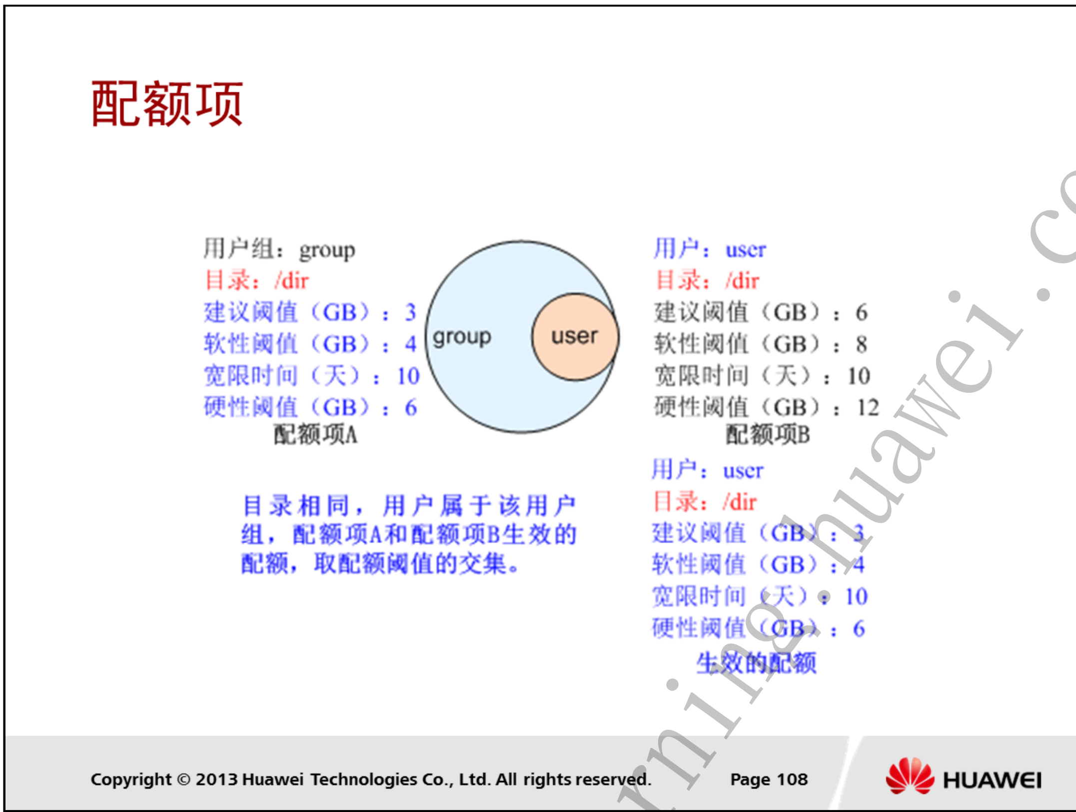
HUAWEI

InfoQuota类型

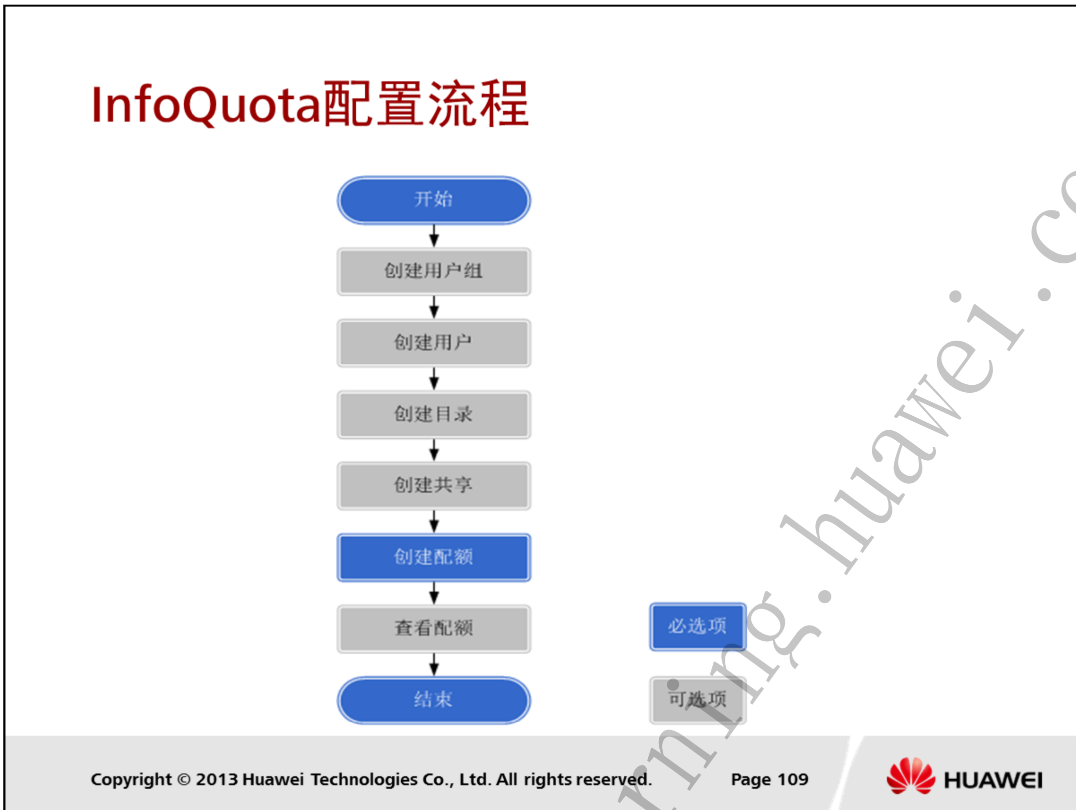
- 配额类型
  - 计算配额：只监控存储空间使用情况不限制使用
  - 强制配额：监控存储空间使用情况的同时限制存储空间的使用
- 阀值类型

阈值类型	说明
建议阈值	当存储空间到达建议阈值时，OceanStor 9000不会限制数据写入，仅上报告警信息。
软性阈值	当存储空间到达软性阈值时，在宽限时间内OceanStor 9000继续允许数据写入，超期后OceanStor 9000立即限制数据写入，同时上报告警信息。 配置软性阈值时，可同时配置宽限时间。
硬性阈值	当存储空间达到硬性阈值时，OceanStor 9000立即限制数据写入，同时上报告警信息。

HUAWEI



- 创建配额项时，需要同时选择用户/用户组和目录，即用户/用户组与目录路径唯一确定一个配额项。
- 当创建的多个配额项的目录或用户/用户组存在交集或子集时，最终生效配额取配额阈值的交集
- 配额控制可以控制文件逻辑存储容量
- 配额管理统计文件的逻辑容量的总和，不包括文件实际的物理存储容量开销和元数据开销
- 可创建的最大配额项8192个
- 只允许对空目录设置配额
- 创建配额的目录深度最多为8级
- 选择用户/用户组和目录，即用户/用户组与目录路径唯一确定一个配额项
- 多个配额项存在交集或子集时，生效配额取配额值小的



- 已经创建了待创建配额的目录
- 已经创建了待创建配额的本地认证或AD域用户/用户组
- 待创建配额的目录为空目录
- 创建配额的目录深度最多为8级
- 已创建的配额项未达到最大规格

配额类型和阀值类型

参数名称	参数说明
配额路径	• 待创建配额的目录。 • 可以直接输入待创建配额的目录名称，也可以单击“选择”，从已经创建的目录中选择配额路径。
位置	• 待创建配额的用户/用户组位置
用户/用户组	• 待创建配额的用户/用户组名称。 • 查找后选择已经创建的用户/用户组。
配额	1. 计算配额 2. 强制配额建议阈值 3. 软性阈值 三个阈值大小依次递增，最小阈值单位为GB。

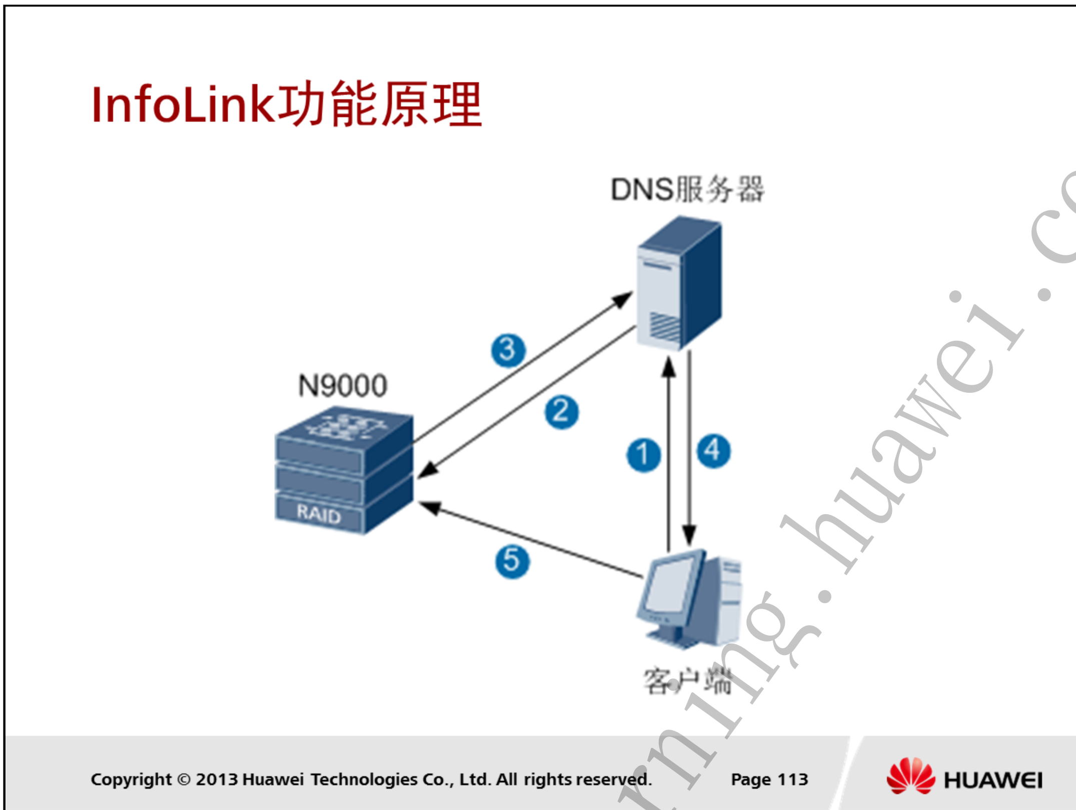
 Page 110 HUAWEI

- 计算配额：只监控存储空间使用情况不限制使用。
- 强制配额：监控存储空间使用情况的同时限制存储空间的使用。
- 强制配额阈值包括：
- 建议阈值：当使用的存储空间到达建议阈值时，不会限制数据写入，仅上报告警信息
- 软性阈值：当使用的存储空间到达软性阈值时，上报告警信息，在宽限时间内继续允许数据写入；超期后立即限制数据写入，同时上报告警信息。配置软性阈值时，需要同时配置宽限时间。
- 硬性阈值：当使用的存储空间达到硬性阈值时，OceanStor 9000立即限制数据写入，同时上报告警信息。

目录

InfoLink技术背景

- InfoLink是客户端连接负载均衡特性，用于管理客户端连接到OceanStor 9000。
- 客户端接入OceanStor 9000时，将用户请求发送到相对空闲的节点，保障了客户端的访问效率
- 当节点出现故障时，使用故障切换功能，将连接到故障节点的业务均衡的分散到其他节点，提高了系统可靠性



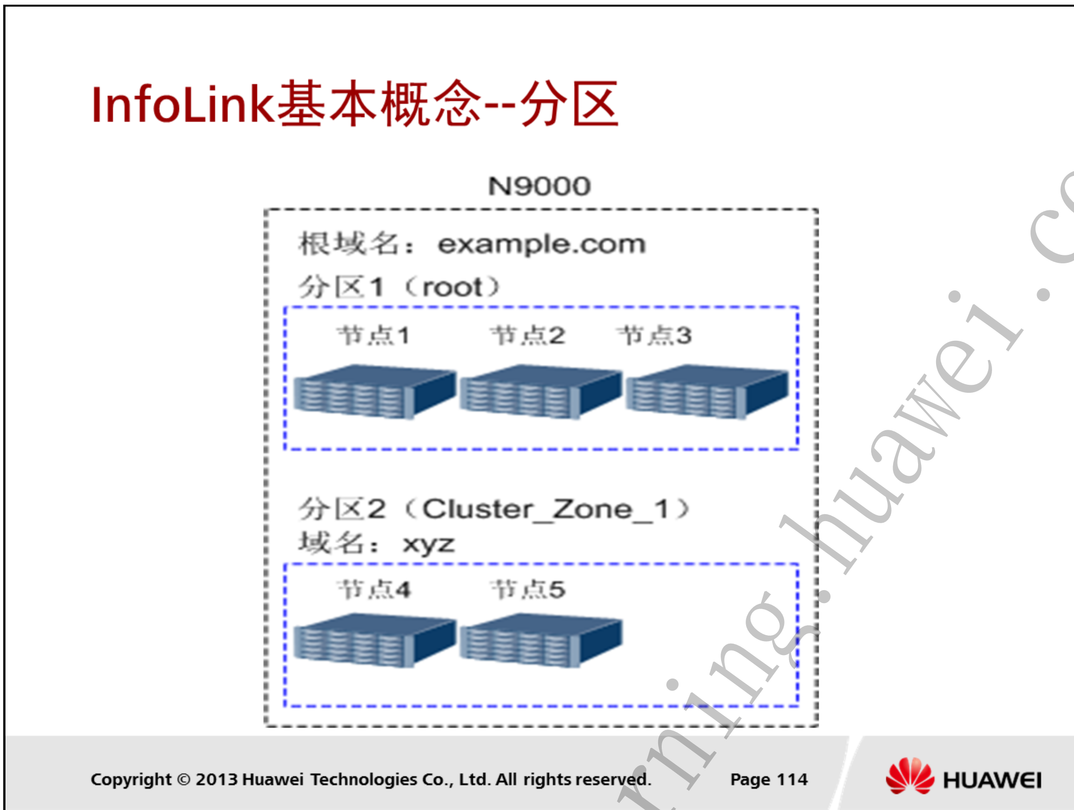
① 客户端使用域名访问OceanStor 9000，发送域名解析请求到DNS服务器

② DNS服务器判断该域名需要由OceanStor 9000解析，将解析请求发送到OceanStor 9000进行解析。DNS服务器转发解析请求的目的IP地址为OceanStor 9000的DNS IP地址

③ OceanStor 9000在域名所对应的分区内部，按照负载均衡策略选择节点，并将该节点的IP地址返回给DNS服务器

④ DNS服务器将节点IP地址返回给客户端

⑤ 客户端使用该IP地址访问OceanStor 9000



假设分区1为root分区，客户端可以使用InfoLink根域名“example.com”接入root分区，也可以使用“default.example.com”接入该分区。分区2为手工创建分区，分区名称为“Cluster_Zone_1”，域名配置为“xyz”，客户端使用域名“xyz.example.com”接入该分区。

多个节点组成的集合，可以为分区配置客户端接入的域名，以及分区内部各节点的负载均衡方式。root分区为默认分区，由系统自动创建。

当新节点加入集群时，将自动分配到root分区。客户端可以使用管理员配置的InfoLink根域名接入到root分区，InfoLink根域名是为集群配置的域名，该域名必须配置为全称域名，即包括根节点在内的域名。

可以手工创建其他分区，将节点划分到指定的分区。需要为各分区配置域名，客户端使用分区域名接入到相应的分区。该域名是InfoLink根域名的下级域名。分区的域名配置为相对域名，自动将InfoLink的根域名作为分区域名的后缀，组成分区的全称域名。最大支持32个分区。

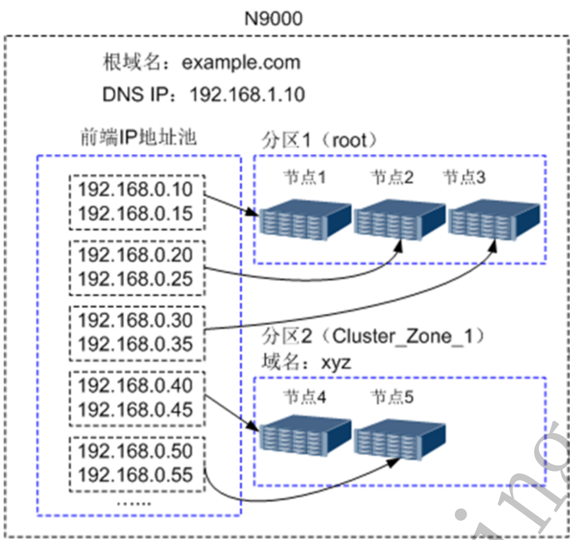
InfoLink负载均衡策略

- 轮询方式
  - 按顺序选择节点处理客户端连接请求
- CPU使用率
  - 选择CPU使用率最低的节点处理客户端连接请求
- 节点连接数
  - 选择客户端连接数量最低的节点处理客户端连接请求
- 节点吞吐量
  - 选择吞吐量最低的节点处理客户端连接请求
- 节点综合负载
  - 按节点综合负载选择节点处理客户端连接请求

 Page 115 HUAWEI

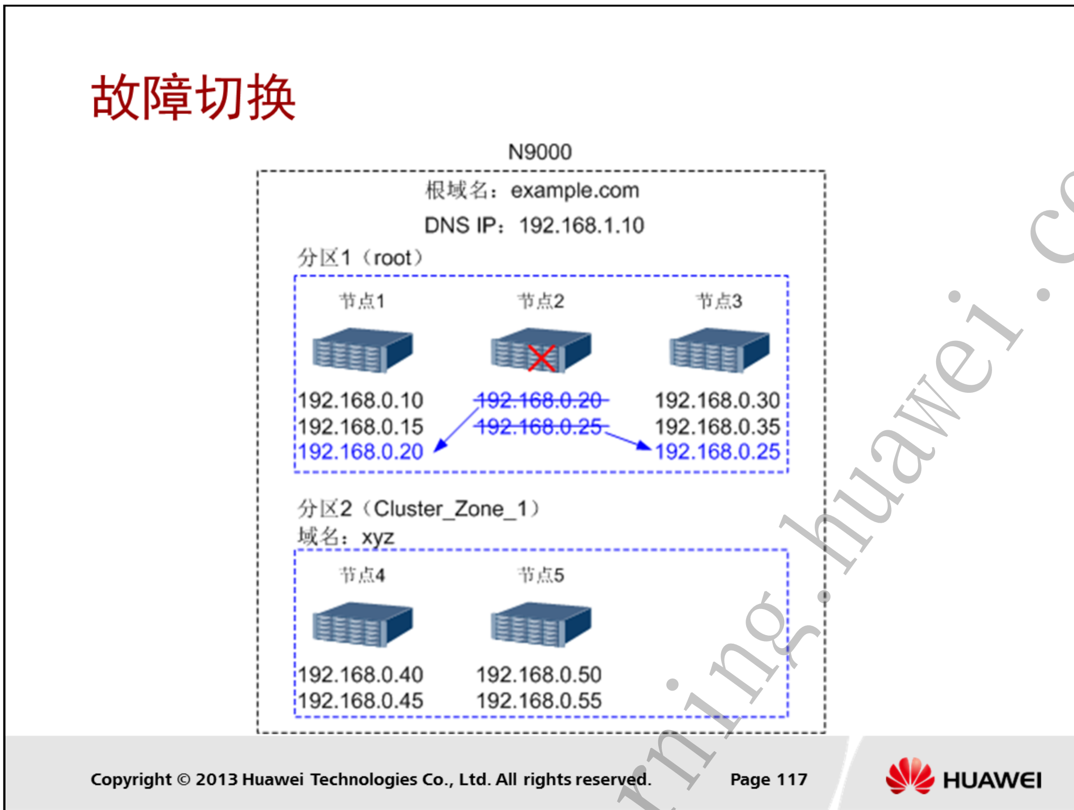
- 根据CPU和网络负载计算节点综合负载，负载越低则被选中的概率越高
- 在未打开license该功能项时，系统使用轮询方式进行负载均衡

智能IP管理



 Page 116 HUAWEI

- 为节点分配静态前端IP地址和动态前端IP地址
  - 静态前端IP地址　该IP地址固定分配给某一个节点
  - 动态前端IP地址　该节点故障后，该IP可以自动分配到其他工作正常的节点
- 配置前端IP地址池，用于为节点分配动态IP地址
- 使用动态前端IP地址实现故障切换功能

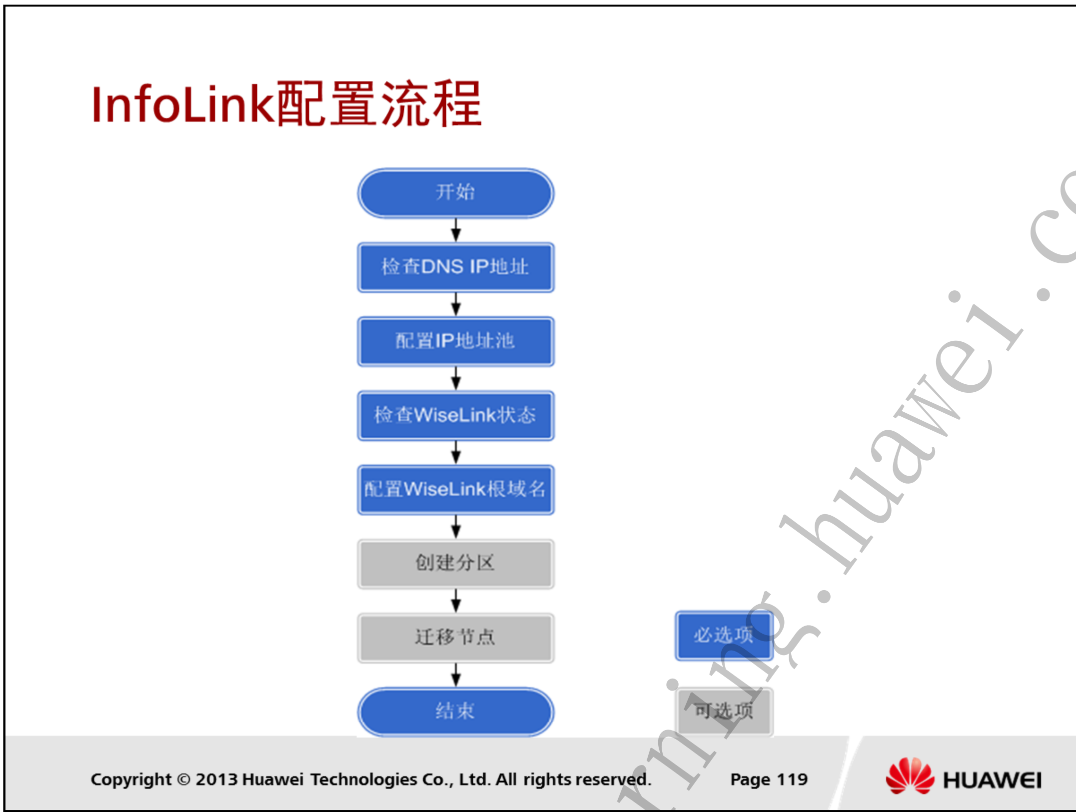


- 节点故障后，为其分配的动态IP地址将被回收，分配给其它正常节点
- 故障节点恢复正常并重新上线后，将为该节点重新分配一个新的动态IP地址
- 如果当前IP地址池中没有空闲的动态IP地址时，将从同一分区中动态IP地址最多的节点迁移动态IP地址到恢复正常的节点，保证该节点正常工作
- 为了避免此类地址切换给在线业务带来影响，建议配置的动态IP地址个数是接入前端网络的接口数量的两倍

故障切换对业务的影响

- CIFS协议
  - 客户端会记录存储系统的状态，因此发生故障切换后将导致连接中断，客户端需要再次连接存储系统
- NFS协议
  - NFS V3协议　客户端每次访问存储系统时将自动重新连接存储系统，因此客户端的连接不会受到影响
  - NFS V4协议　与CIFS协议类似，客户端将受到影响，需要客户端重连

HUAWEI



1. 检查DNS IP地址：InfoLink特性需要OceanStor 9000向客户端提供域名解析服务，因此需要配置DNS代理IP地址。在部署OceanStor 9000时，已配置该IP地址，此时需要检查IP地址是否正确。
2. 配置IP地址池：InfoLink特性中，节点将通过IP地址池中的IP地址和客户端通信。添加IP地址后，该IP地址会自动分配给节点。
3. 检查InfoLink状态：在部署OceanStor 9000时默认已开启InfoLink，此时需要确认InfoLink已开启。
4. 配置InfoLink根域名: 客户端可以使用根域名接入OceanStor 9000。
5. 创建分区:为了便于管理，管理员可以创建多个分区。每个分区可以作为独立的管理单元，配置负载均衡策略和域名。
6. 迁移节点:当管理员创建了多个分区时，可以将root分区下的节点迁移到新创建的分区中。

总结

- 大数据存储硬件组成和软件特性
- OceanStor 9000系统典型组网
- OceanStor 9000软硬件安装
- NFS和CIFS配置部署
- Info系列特性原理与配置

思考题

- OceanStor 9000硬件节点有哪几种类型？请描述下节点硬件组成
- OceanStor 9000典型组网有哪几种？请描述下组网形态和应用场景
- 请描述一下9000系统初始化配置和文件系统共享的基本配置过程？

习题

练习题：

1、（选择题）9000系统中NFS协议支持的认证方式有：（）

A、LDAP；　B、NIS；　C、客户端认证；　D、AD域认证

2、（选择题）9000 V100R001版本组网支持以下哪些?

A 、前端10GE后端IB； B、 前端10GE后端10GE； C、 前端IB后端IB；

D 、前端IB后端10GE

3、（判断题）9000所有节点都可以提供NFS/CIFS服务?

HUAWEI

- 1 AB；2 ABCD；3、T

Thank you

www.huawei.com

HC1209205 集群 NAS 存储系统故障处理

HC1209205
集群NAS存储系统故障处理

www.huawei.com

HUAWEI

目标

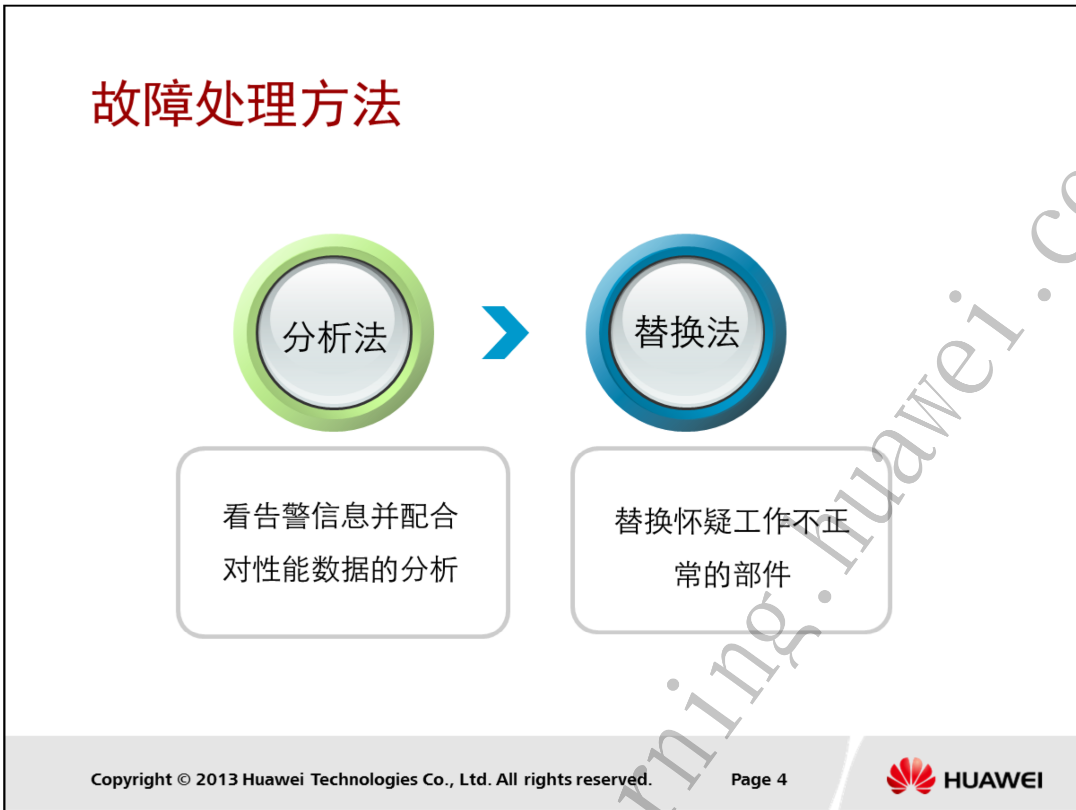
- 学完本课程后，您将能够:
  - 了解N8500问题处理流程
  - 掌握N8500常见问题及故障的处理方法
  - 具备集群NAS典型故障分析处理能力

目录

- N8500故障处理的原则、方法
- N8500故障处理的流程
- N8500分类故障以及典型问题处理方法
- N8500故障处理案例



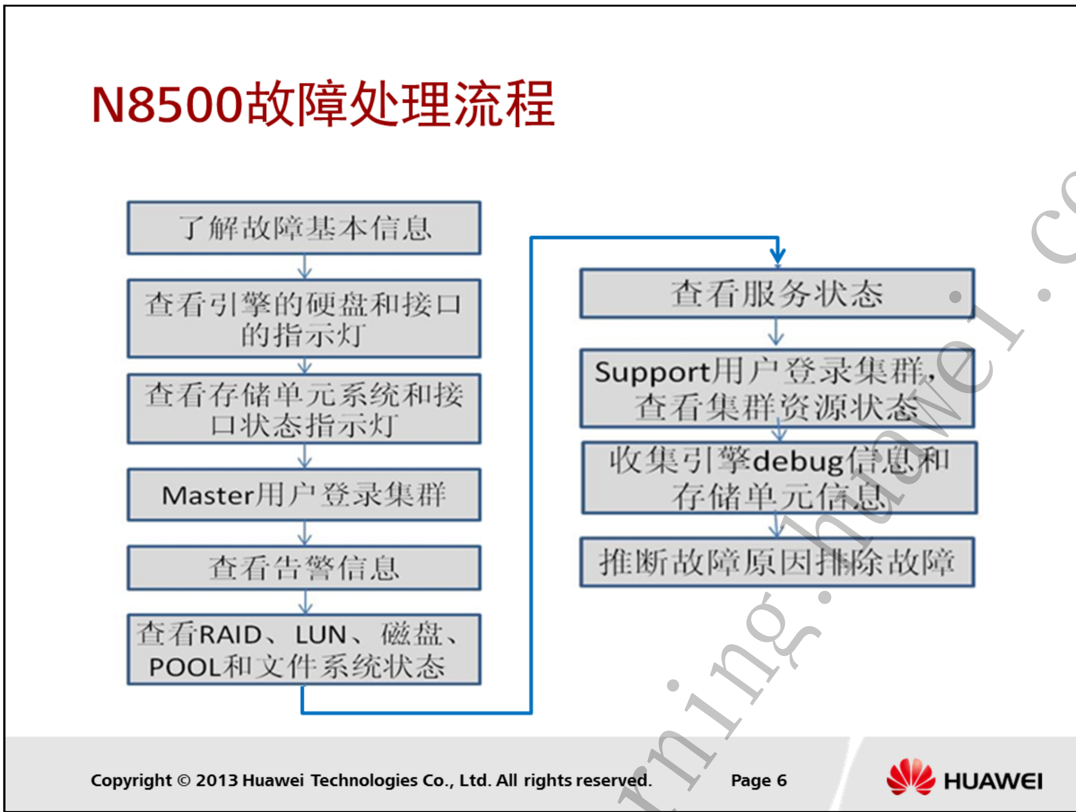
- 故障处理的原则：
  - 首先排除外部设备的问题。外部设备问题包括光纤、网线、客户设备和掉电等问题；内部问题包括产品部件的故障问题。
  - 在分析告警时，应该按照级别的优先级依次进行分析，如先分析紧急级别的告警，再分析重要和警告级别的告警。
  - 在分析告警时，应先分析是个别问题还是共性问题，确定问题的影响范围。需要确定是一个模块出问题，还是多个模块出现类似问题。



- 故障处理方法
  - 当系统发生故障时，一般会伴随有大量的告警信息产生，通过查看告警信息并配合对性能数据的分析，可大概判断出所发生故障的类型和位置。
  - 替换法就是使用一个工作正常的部件去替换一个怀疑工作不正常的部件，从而达到定位故障、排除故障的目的。这里的部件可以是一段光纤跳线、一根网线、一个电源模块或者一个风扇模块。

目录

- N8500故障处理的原则、方法
- **N8500故障处理的流程**
- N8500分类故障以及典型问题处理方法
- N8500故障处理案例


N8500故障处理流程
了解故障基本信息
查看引擎的硬盘和接口的指示灯
查看存储单元系统和接口状态指示灯
Master用户登录集群
查看告警信息
查看RAID、LUN、磁盘、POOL和文件系统状态
查看服务状态
Support用户登录集群，查看集群资源状态
收集引擎debug信息和存储单元信息
推断故障原因排除故障
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 6
HUAWEI



- N8000设备的版本包括：
  - 集群NAS引擎软件的版本
  - 存储单元的软硬件版本
  - ISM管理软件的版本
- 收集N8000设备的日志信息包括：
  - 集群NAS引擎的debug信息
  - 存储单元的日志信息
  - ISM管理软件的日志信息

集群NAS引擎日志系统

日志位置	日志文件	日志用途
/var/log	message	操作系统相关的所有关键事件
/var/log	sfsfs_event.log	引擎的network，storage服务所产生的事件日志
/opt/VRTSnasgw/log	Upgrade.log	引擎系统升级中的过程
/opt/VRTSnasgw/log	scanbus.log	引擎扫描磁盘操作中搜集的相关信息和关键事件
/opt/VRTSnasgw/log	CIFS.log	CIFS共享操作相关的集群管理服务操作及相关事件
/opt/VRTSnasgw/log	NFSAgent.log	NFS共享代理操作相关的集群管理服务操作及相关事件
/var/VRTSvcs/log	Engine_A.log	VCS集群日志

  HUAWEI

故障分析定位和排除

- 故障分析判断和定位
  - 通过对收集的信息进行分析，从众多可能原因中找出故障原因的过程，通过故障分析，可以确定故障范围、故障种类、故障发生的具体原因以及故障排除的手段。
- 故障排除
  - 是指采取适当的措施或步骤清除故障、恢复系统及业务的过程，具体的方式有检修线路、更换硬件、修改配置数据、重新启动服务或应用程序、替换损坏文件、重新启动系统、修复文件系统及服务等

HUAWEI

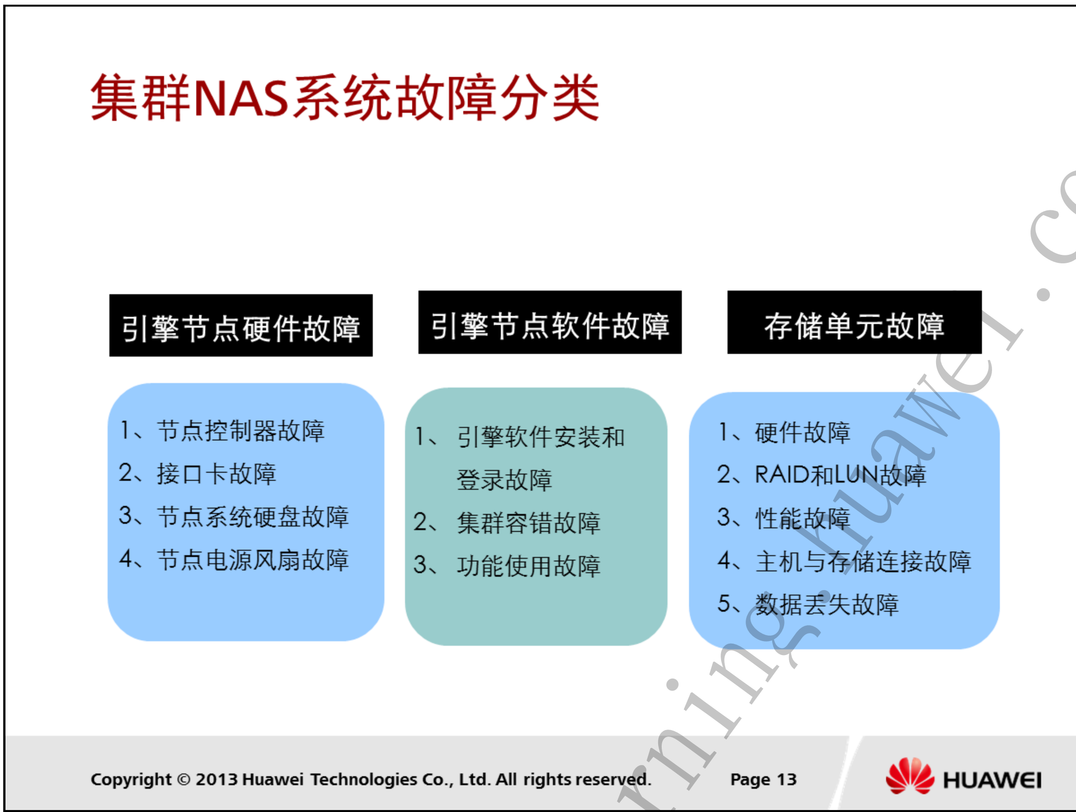


- 原始信息分析
    - 主要用来判断故障的范围、确定故障种类，原始信息分析可以为缩小故障的定位范围和初步定位问题提供依据
- 指示灯状态分析
    - 系统硬件（如集群系统、存储系统）都提供了相应的运行、状态和特性指示灯，可以直接反映硬件的工作状况，是故障分析和定位的重要依据之一
    - 指示灯状态分析用于快速查找硬件故障原因，由于指示灯所包含的信息量相对不足，应与告警信息分析配合执行
- 告警信息分析
    - 告警信息是指系统输出的告警信息，通常以指示灯、声音、屏幕输出等形式提供给维护人员，具有简单、明了的特点
    - 告警信息包含故障现象、可能引起故障的原因、修复建议等，是故障分析的重要依据之一，可通过告警信息定位故障发生的位置和大致原因



- 告警信息分析
  - 告警信息是指系统输出的告警信息，通常以指示灯、声音、屏幕输出等形式提供给维护人员，具有简单、明了的特点
  - 告警信息包含故障现象、可能引起故障的原因、修复建议等，是故障分析的重要依据之一，可通过告警信息定位故障发生的位置和大致原因
- 日志信息分析
  - 本系统提供详细的日志记录功能，故障分析时可通过日志信息了解系统运行状况。相对于告警信息，日志信息更详细和具体，可直接用于定位故障发生的原因
- 消息跟踪分析
  - 消息跟踪可用于分析上网流程和设备间消息交互流程。利用跟踪结果，维护人员可直接定位故障原因，或从中得到启发，为后续分析提供宝贵的思路

目录



- N8000引擎节点硬件故障主要包括：
  - 节点控制器的硬件故障
  - 节点接口卡的故障，包括业务端口接口卡故障和连接后端存储单元的接口卡故障
  - 节点系统硬盘故障
  - 节点电源和风扇故障
- 引擎节点软件故障主要包括：
  - 引擎软件安装和登录故障：无法通过master账号登录、CLI登录异常、引擎之间的信任关系异常
  - 集群容错故障：物理路径异常导致的业务挂死、节点异常导致的集群挂死
  - 功能使用故障：iofencing功能故障、文件系统故障、cifs应用故障、NFS应用故障

集群NAS系统故障影响

引擎节点硬件故障

1、业务中断
2、业务性能下降

引擎节点软件故障

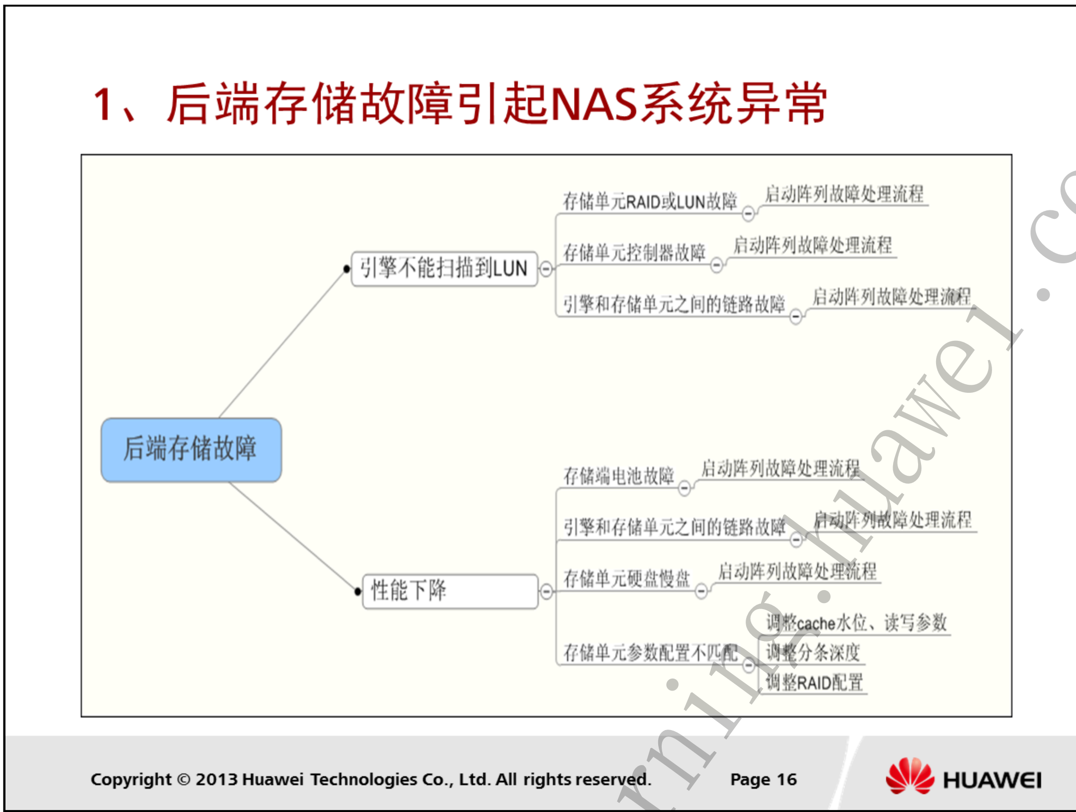
1、 无法登录
2、 集群节点无法切换
3、 业务性能下降
4、业务功能无法使用

存储单元故障

1、文件资源失效
2、业务中断
3、性能下降
4、数据丢失

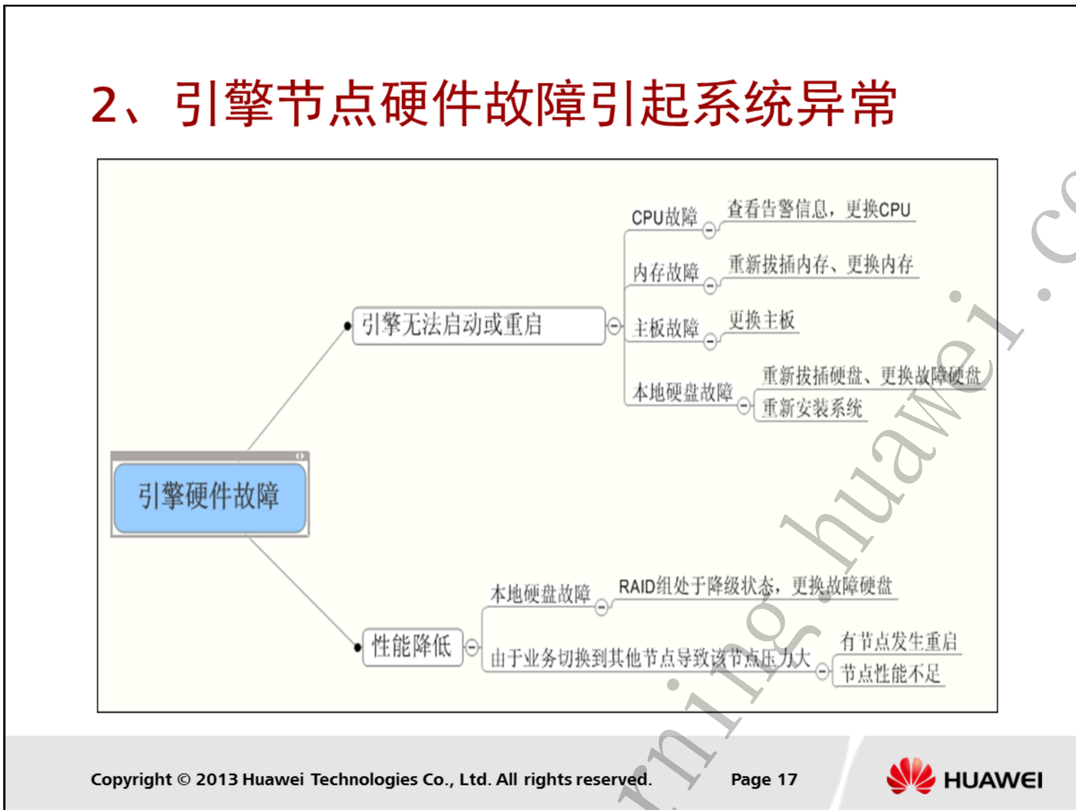
集群NAS典型故障诊断

1. **后端存储故障引起系统异常**
2. 引擎节点硬件故障引起系统异常
3. 内部通信链路问题诊断
4. 管理模块问题诊断
5. 文件系统故障诊断
6. 文件读写性能问题诊断
7. NFS挂载和访问问题诊断
8. CIFS挂载和访问问题诊断
9. 网络问题诊断



1. 判断是否由于后端存储单元的原因导致业务中断，首先需要确认FC网络是否正常，然后确认引擎能否发现磁盘，然后再从存储单元侧发现原因。

2. 如果性能下降首先检查业务网络的性能情况（需要检查业务交换机），其次需要检查文件系统参数配置情况，然后查看FC网络的速率和运行情况，最后查看存储单元LUN的运行情况，包括LUN的参数是否和业务匹配，如果都没问题，需要检查RAID组是否降级或者处于重构状态，如果也没问题，需要检查RAID组中是否存在慢盘。



- 引擎节点的硬件故障问题，可以参考通用的服务器故障诊断和处理流程

  I1910　Intrusion detected. Check chassis cover

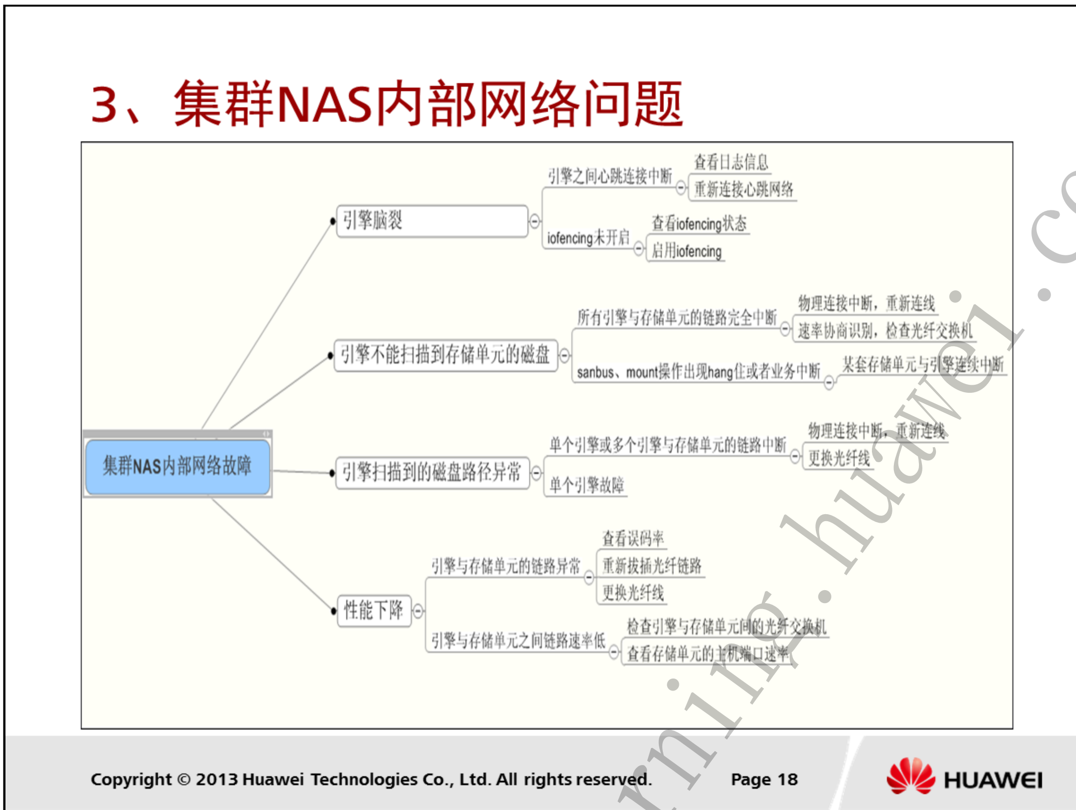
  E1414　CPU # temp exceeding range. Check CPU heatsink

  E2118　Unrecoverable Memory CRC error. Power cycle AC

  E1812　Hard drive ## removed. Check drive

  E1310　Fan ## RPM exceeding range. Check fan

  E1610　Power Supply # (### W) missing. Check power supply



集群NAS的内部网络包括引擎之间的心跳网络、引擎与存储单元间FC网络和存储单元控制框与硬盘框之间的连接。

内部通信链路问题分析和处理方法

- 引擎之间心跳网络故障

序号	问题	解决方案
1	脑裂	1、停止业务 2、修复心跳网络 3、配置iofencing硬盘，引擎识别到iofencing硬盘，但是不用加入pool 4、启用iofencing功能，重启整个集群

HUAWEI

supoort账户登录节点， 输入vxdctl –c mode ，如果显示每个节点都是主节点，表明已经发生脑裂，需要修复心跳，并启用iofencing。如果发生脑裂，需要立即停止业务，否则将会导致数据异常。

内部通信链路问题分析和处理方法

- 引擎与存储单元的链路故障

序号	问题	解决方案
1	某个节点不能扫描到硬盘	1、检查该节点的系统状态 2、查看该节点的HBA卡的状态并恢复正常
2	执行scanbus、mount操作时挂死	1、检查所有存储单元的状态 2、查看所有存储单元与引擎的FC连接状态，如果有存储单元的FC连接异常，需要修复连接。 3、在引擎主节点的support模式下，通过kill命令删除scanbus进程 4、重启主节点

HUAWEI

- 引擎在执行scanbus扫描磁盘后，磁盘在某些节点上面异常，有可能某个引擎与存储单元的FC链路异常。
- 在有多套存储单元时，如果执行scanbus、mount操作，节点长时间没有反应，有可能某个存储单元与所有引擎的FC链路异常。

内部通信链路问题分析和处理办法

序号	问题	解决方案
1	存储单元端FC误码率过高	1、更换光纤线、光模块和FC主机口 2、更换控制器
2	引擎、存储单元与光纤交换机的协商问题	1、 登录存储单元，查看HBA卡的连接状态和速率，修改HBA的连接状态为点对点或者交换机模式 2、登录光纤交换机，更改端口模式

Page 21
HUAWEI

- 如果引擎端和存储单元侧都没有发现明显的问题，可以查看FC光纤网络的误码率和光纤线的质量情况。
- 如果引擎和存储单元之间通过光纤交换机连接，可以检查引擎、光纤交换机和存储单元之间的速率协商情况。
- 如果是硬盘框或者硬盘框的与控制框连接故障，可以参考表格中第2条处理。

4、管理模块问题

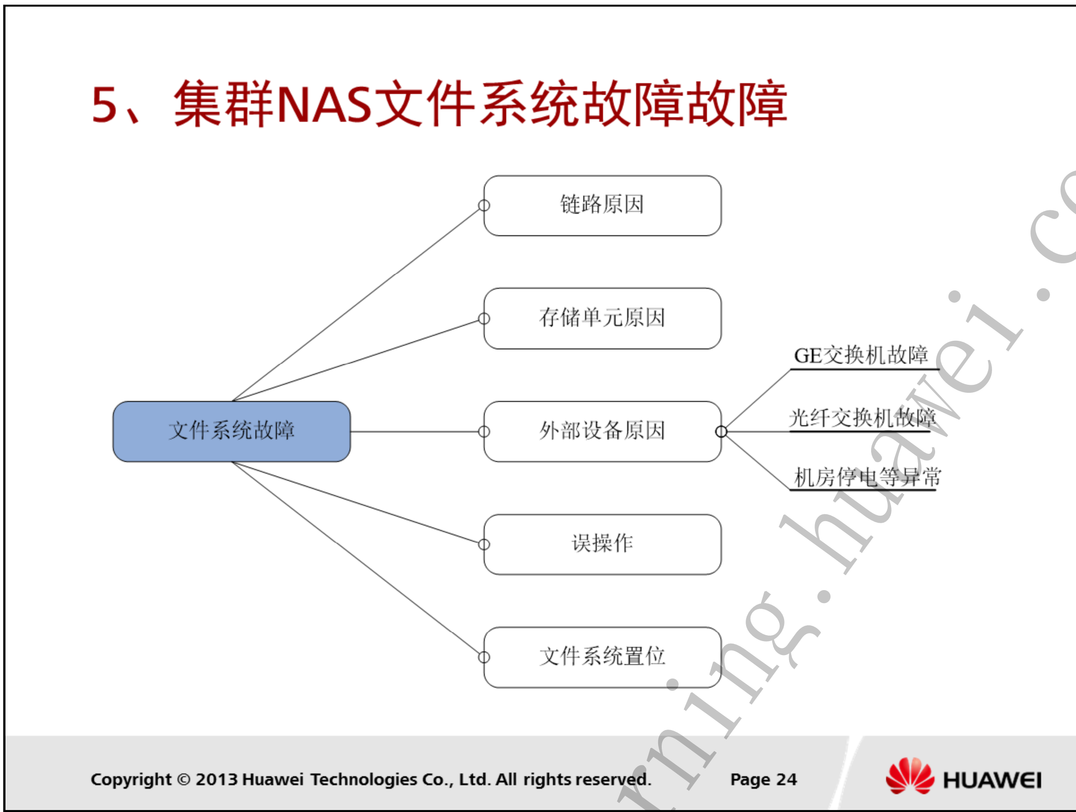


HUAWEI

管理模块问题解决方案

序号	问题	解决方案
1	console口所在的物理网卡没有连网线	连上网线，等待1分钟后重新登录
2	console地址与网络上其他主机有冲突	1、通过KVM修改console地址 2、修改冲突主机上的IP地址
3	资源offline	1、通过hagrp -stat | grep Man 2、通过hagrp -online ManagementConsole -sys N8300_01 命令上线

HUAWEI



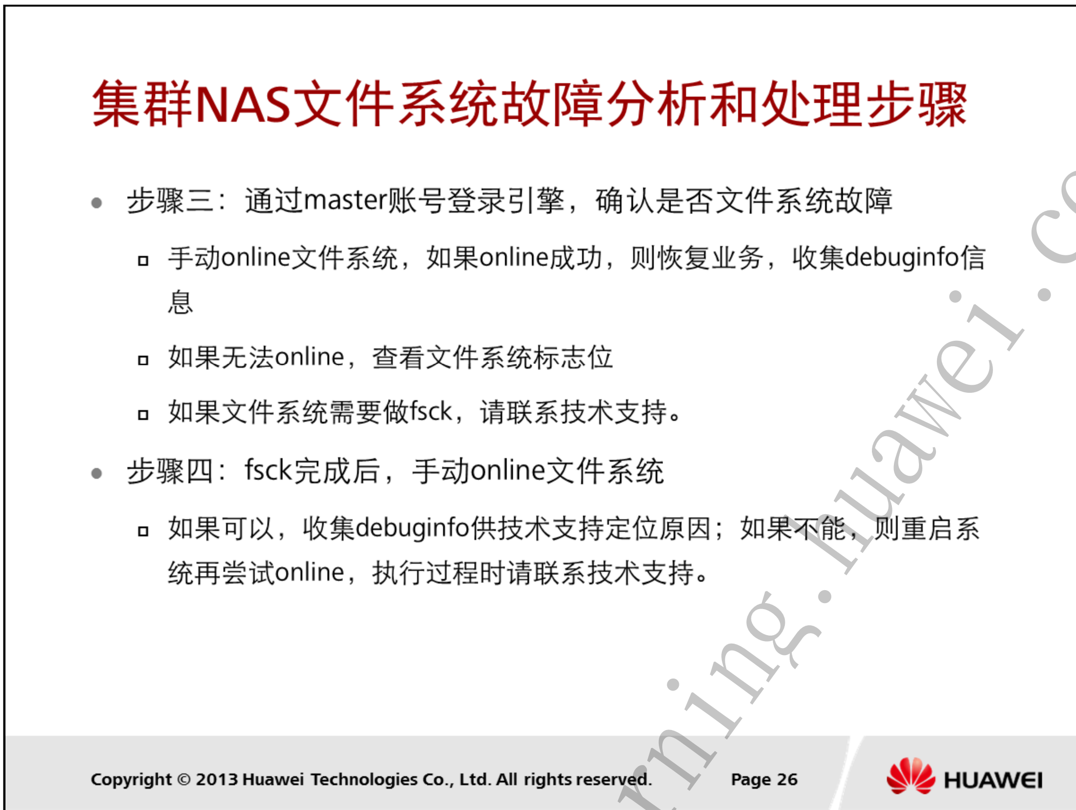
- 文件系统故障处理原则:
  - 向客户确认故障现象、时间、故障前执行的操作、是否出现继发掉电、交换机故障等事故;
  - 按照链路、存储单元、文件系统的顺序逐步排查;
  - 出现RAID失效、误删LUN、文件系统需要fsck等重大故障，需要联系技术支持人员
  - 处理完成后，务必收集debuginfo供定位根因。

集群NAS文件系统故障分析和处理步骤

- 步骤一：确认引擎和存储单元间的链路是否正常
  - 登录N8000，执行storage disk list paths，查看每个节点链路状态
  - 若存在非active的链路，恢复链路后执行storage scanbus，手动online文件系统
- 步骤二：登录存储单元，确认存储单元状态
  - 查看是否有存储单元故障，参考存储单元故障处理内容。
  - 是否误删N8000文件系统在用的LUN，请联系技术支持
  - 是否误删映射，恢复LUN映射，执行storage scanbus，文件系统可自动恢复，若无法恢复，联系技术支持。

 Page 25 HUAWEI

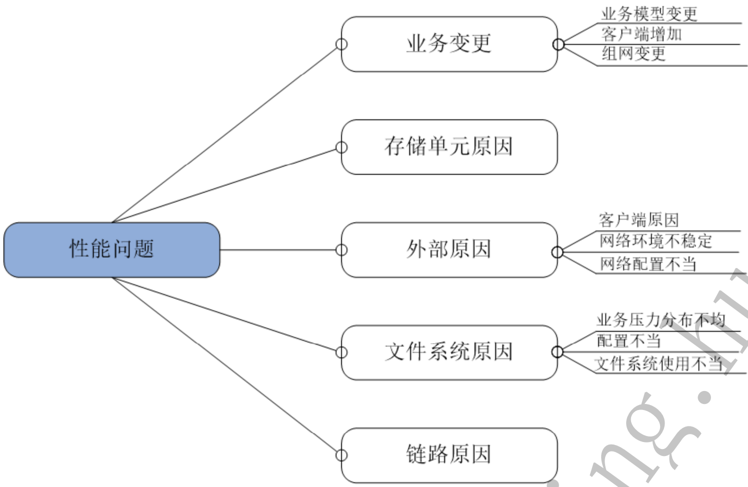
注：上述storage disk list paths，storage scanbus，命令均通过master账户登录到N8000命令行界面执行。



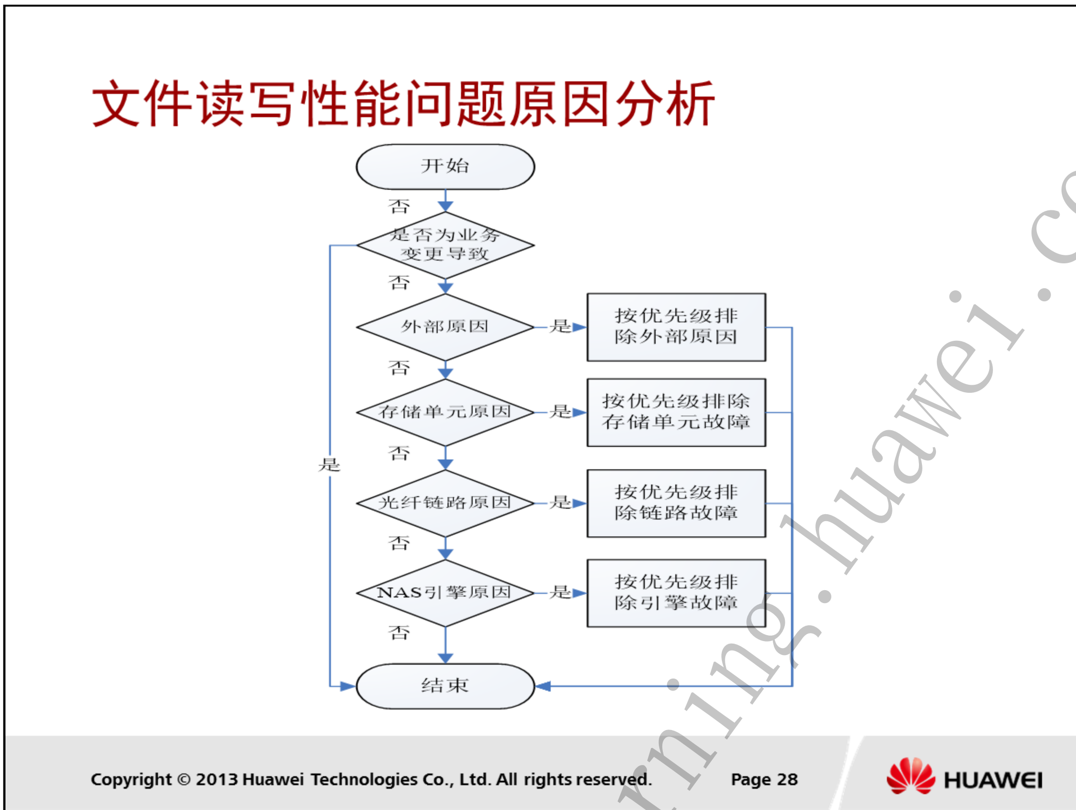
查看文件系统标志位时，使用support账户登录N8000，执行echo “8192B.p S” | fsdb -t vxfs ; /dev/vx/rdsk/sfsdg/sfsvol|grep flag，其中sfsvol为文件系统名，可能的显示为flags 0 mod 0 clean 3c或flags 4000 mod 0 clean 3c，如果flags最后一位出现1，则文件系统被置位，需要使用fsck命令进行修复。

注：手动online文件系统时，使用master账户登录到N8000命令行界面，执行storage fs online fsname，其中fsname为故障文件系统名。命令成功返回后，使用storage fs list命令确认该文件系统状态为online。

6、文件读写性能问题



HUAWEI

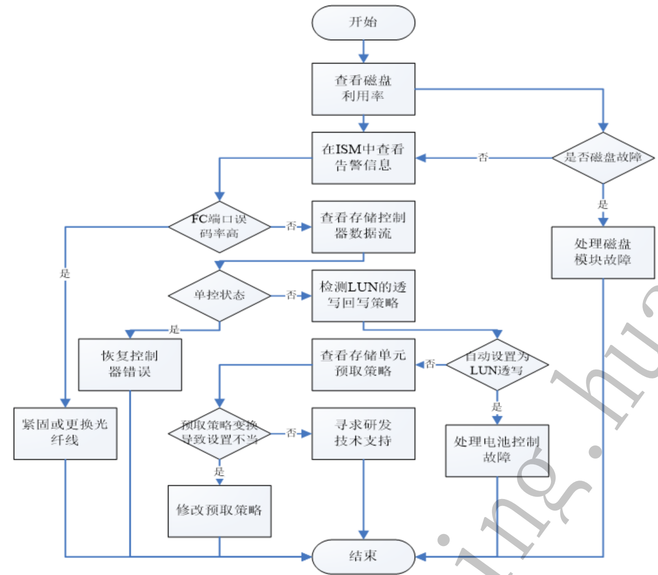


- 文件系统性能下降时，应按照从外到内、从下到上的顺序进行逐一排查：
  - 业务变更原因
  - 外部原因
  - 存储单元原因
  - 链路原因
  - 集群NAS引擎原因

文件读写性能问题诊断-业务变更

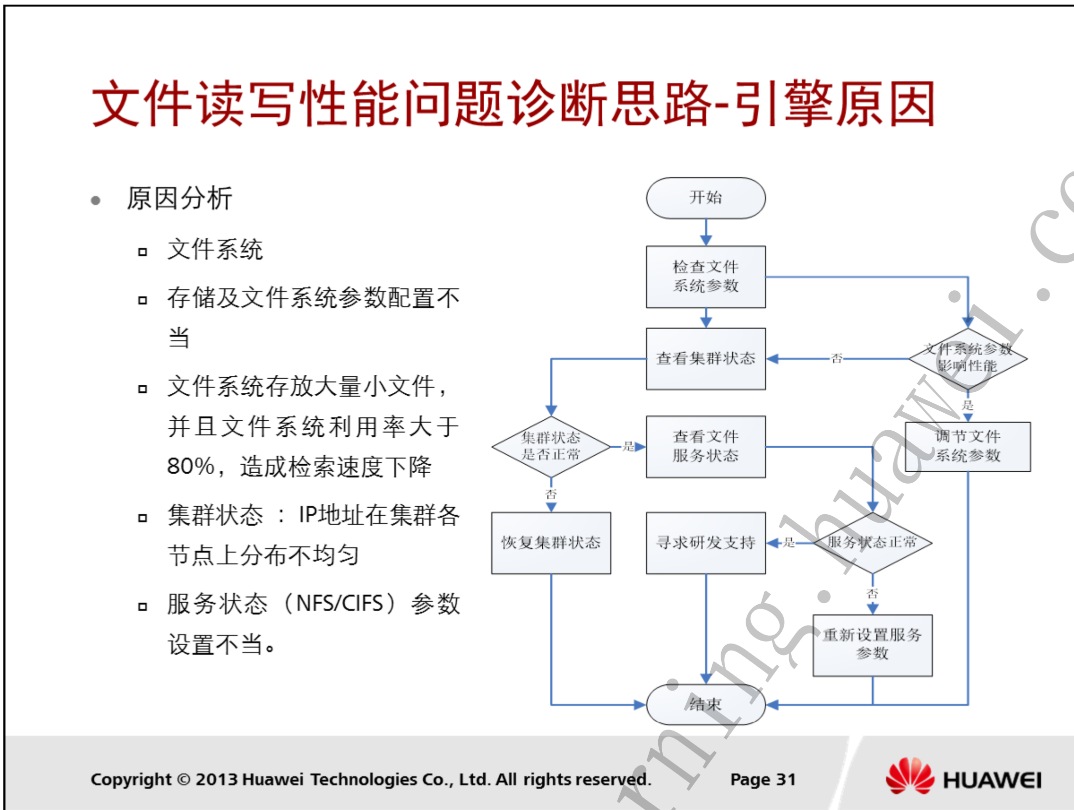
- 原因分析
  - 原来只有顺序读写业务，现在变成了随机读写业务；
  - 原来有只有读业务，现在加入了写业务；
  - 原来有10路用户并发，现在更多路用户同时并发；
  - 其他复杂业务变更。
- 处理意见：
  - 由于业务类型的变化而造成的流量上的下降属正常现象，无需处理。

文件读写性能问题诊断-存储单元原因



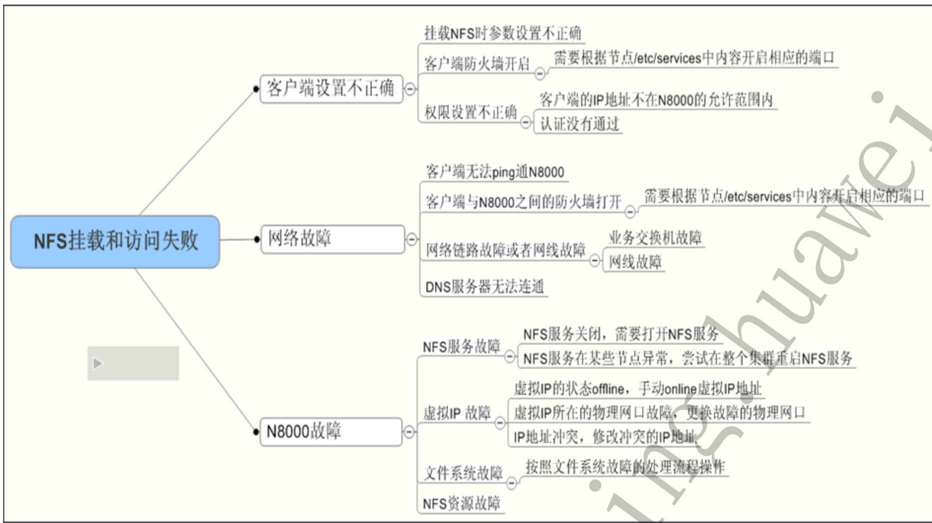
 Page 30 HUAWEI

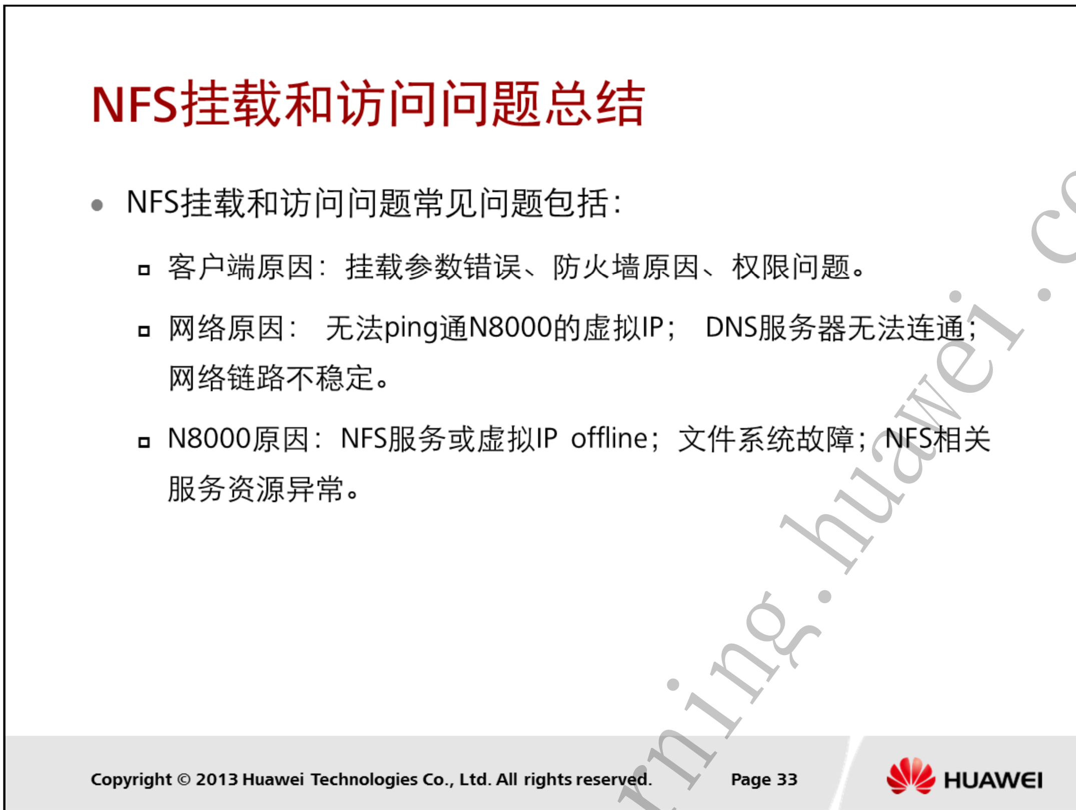
- 原因分析
  - 磁盘故障：出现慢盘、磁盘出现坏道
  - 光模块故障：FC端口误码率很高
  - 控制器故障
  - LUN变透写造成写性能下降
  - 预取策略变化造成读性能下降



- 存储单元配置建议：
  - SAS、FC盘配置9盘、11盘RAID5，9盘RAID条带深度128KB，11盘64KB
  - SATA、NL-SAS盘配置10盘或12盘RAID6，10盘RAID条带深度128KB，12盘64KB
  - RAID组不跨框，保险箱盘单独创建RAID
  - LUN尽量不要超过16TB，写设置-采用回写镜像，预期策略-采取智能预取
  - Lun归属控制配置，a）系统只创建一个RAID组，SATA盘：该RAID组的LUN都归属于一边控制器；SAS/FC盘：一半归属于A控，另一半归属于B控。b）系统创建的RAID组多于1个，同一RAID组的LUN归属于同一控制器，不同的RAID组交替归属于A控和B控。
- 文件系统配置建议：
  - 条带文件系统columns 数量不超过8，防止数据过于离散
  - tripeunit 选择2048，blocksize选择8192
  - 每个Pool对应一个RAID组，即一个Pool里面的disk由一个raid组的lun组成
  - 创建条带/镜像条带/条带镜像文件系统，columns数量与raid组个数一致
  - 在实际配置时，从可靠性和故障恢复速度方面考虑，文件系统大小建议不超过64TB。具体视文件大小决定，对小文件（平均文件大小<1M）存储应用，不超过16TB，文件越小，建议创建越小容量的文件系统。
  - 建议单个文件系统存放文件不超过1亿，目录层度不超过10层，每个目录下不超过10万个文件或目录。

7、NFS挂载和访问问题

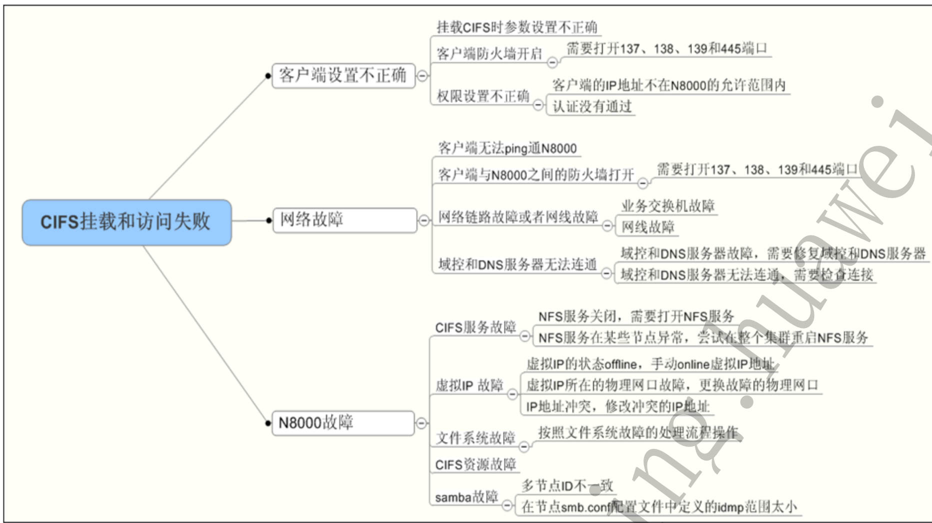


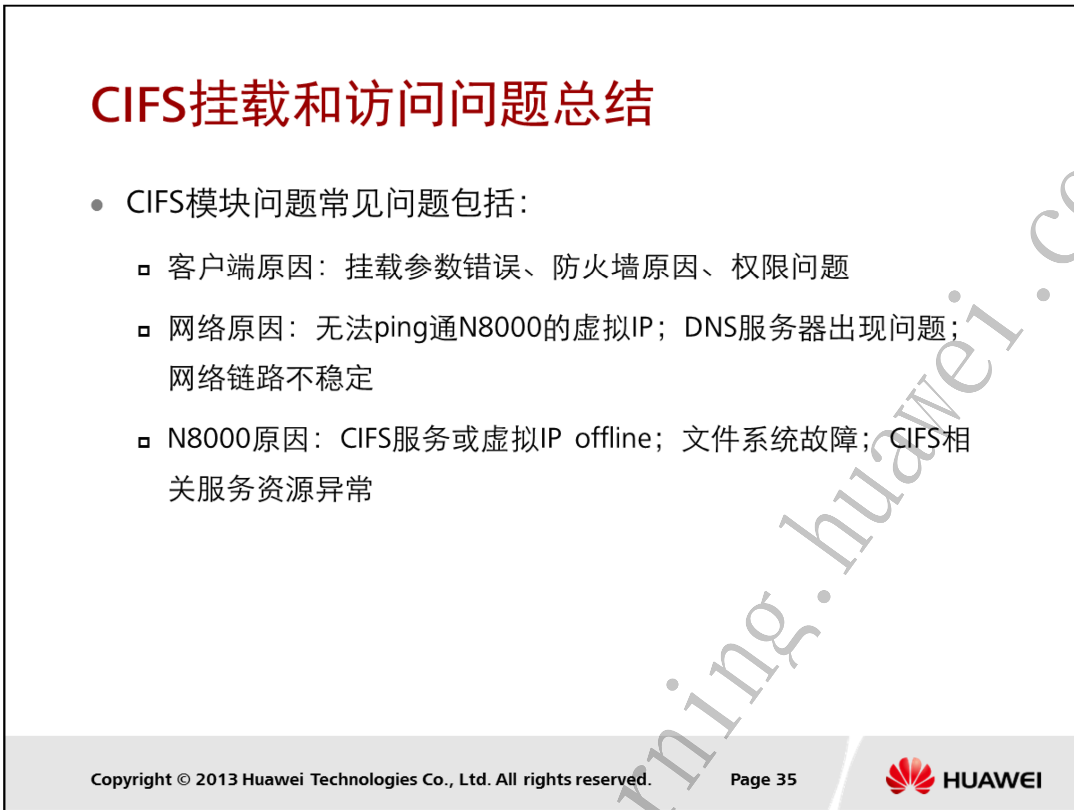


当客户端无法挂载nfs共享时，确认N8000的防护墙是否关闭，可执行iptables －F关闭之。同时在客户端使用showmount －e IP，查看N8000的nfs共享，确认该客户端是否在共享允许的权限范围内。

另外，查看NFS相关服务状态的方法为：Support登录N8000，执行命令hagrp －state | grep NFS，查看是否出现offline或者faulted状态。

8、CIFS挂载和访问问题诊断





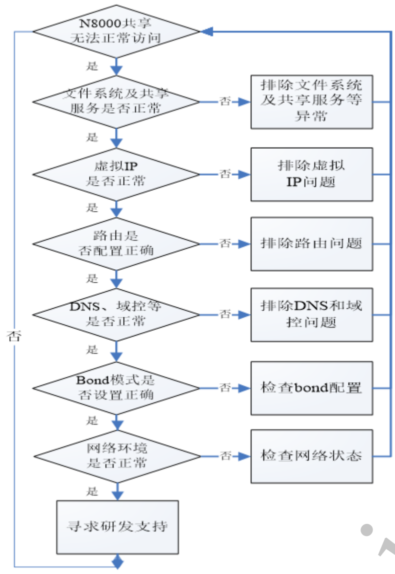
CIFS应用分为Normal和homedir两类，在问题处理前首先需要确认故障种类。

9、NAS网络问题



HUAWEI

NAS网络问题诊断总结



Page 37
HUAWEI

- 网络问题比较复杂，需要逐一排除：
  - 文件系统、共享服务是否正常
  - 虚拟IP是否正常（是否有offline\faulted状态）
  - 路由是否正确配置
  - DNS、域控是否正常或正确配置
  - Bond模式是否正确配置
  - 网络环境是否存在干扰（丢包、乱序、重传等）

目录

HUAWEI

案例1：管理模块无法登录

- 问题描述：
  - 客户反映无法通过master登录集群，显示该账号不可用
- 原因分析：

  1、通过终端可以ping通管理console地址

  2、通过KVM连接到集群，通过support账号登录集群，使用 hastatus –sum | grep Man 查询状态为offline，执行hagrp -online ManagementConsole -sys testN8300_01

  3、在集群的support账号下，使用ethtool查看管理网口的物理状态。通过在主机端检查arp –a 发现consoleip的mac地址不是集群管理网口的mac地址，表明网络中的ip地址有冲突

 Page 39 HUAWEI

- 解决步骤：在网络管理人员的协助下，找到并修改地址冲突的那台主机的IP地址。
- 验证恢复：重新登录集群，通过master用户可以正常登录集群。

案例2

- 问题及故障描述
  - 集群在设置IO fencing后，重启整个集群，出现部分节点无法加入集群的情况，此现象概率出现
- 原因分析
  - 集群重启时，IO防护为了防止集群脑裂，会将部分启动较慢的节点排除出集群。因此不允许直接重启整个集群。若必须将所有节点都进行重启，请一次只重启一个节点，待该节点完全启动并可以提供业务时，再重启下一个节点，直到所有节点全部重启完毕，且可以正常提供业务
- 处理步骤
  - 该现象发生后，重启未能加入集群的节点即可修复

 Page 40 HUAWEI

案例3

- 问题及故障描述
  - 在有NFS业务时，重启一个业务节点后，一个文件系统offline，导致该文件系统不可访问
- 原因分析
  - 重启节点与主节点的时间不一致导致，集群为保护文件系统一致性，强制将文件系统标志位设置为1，引起文件系统offline
- 处理步骤
  - 登录到集群管理界面的storage模式下，使用fs fsck fs100g命令进行文件系统修复，修复完成后需要使用fs online fs100g的命令将文件系统上线

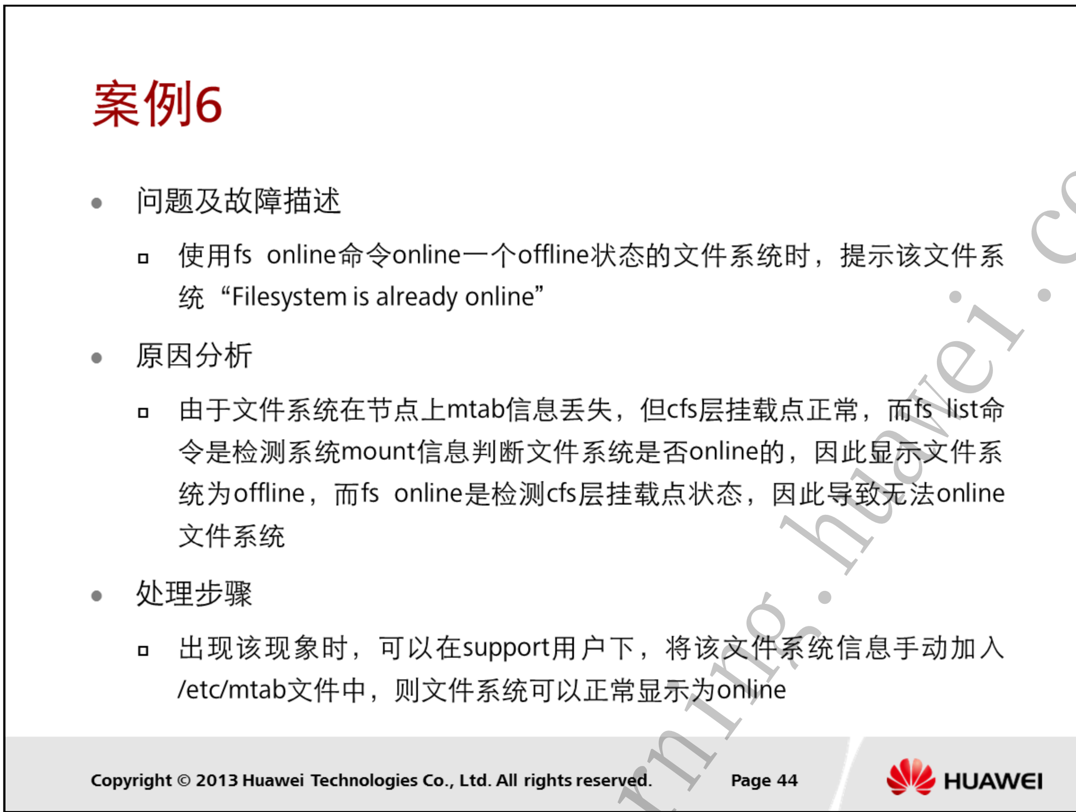
- 如果fs fsck 命令无法成功执行，需要使用support账户登入集群节点，首先使用cfsmntadm display命令查看文件系统是否依然mount，如果依然mount，使用cfsumount命令卸载文件系统
- 在suport账户下，使用fsck –t vxfs –o full –y /dev/vx/dsk/sfsdg/fs100g命令对文件系统做全校验
- 使用fs online fs100g命令上线文件系统

案例4

- 问题及故障描述
  - 把一个节点从集群删除后再加入集群，该节点的NFS共享服务fault，业务IP无法切换到该节点
- 原因分析
  - 在添加一个新的节点加入集群时，必须保证该节点是全新安装的节点，否则由于原来删除的节点依然保留有原有集群的信息，可能导致发生异常情况
- 处理步骤
  - 出现此问题时，执行nfs server start命令重启NFS服务，即可将虚拟IP online到新加入的节点



- 设置时区时先使用clock regions命令查找到设备所在地所属的时区，软后使用clock timezone clock Easten 设置时区即可



- 若依然无法恢复，观察故障文件系统上是否仍有业务运行，若有则先将业务中止，随后在CLI下执行fs offline fsname操作将文件系统下线；接着再执行fs online fsname操作将文件系统上线，即可解决。

案例7

- 问题及故障描述
  - 对未创建快照的文件系统进行删除时，提示有快照存在不可删除
- 原因分析
  - 删除文件系统未成功时会报出此提示，并不是真的有快照未被删除，错误提示信息有误
- 处理步骤
  - 尝试再次执行fs destroy命令删除文件系统
  - 如果无法删除则使用support帐户进入调测模式，使用如下命令删除文件系统
  - 使用cfsmntadm display命令查看文件系统对应的挂载点是否存在，如果存在则使用cfsumount /vx/文件系统名称 命令卸载映射
  - 使用cfsmntadm delete /vx/文件系统名称命令删除挂载点

HUAWEI

案例8

- 问题及故障描述
  - 集群增加节点后，客户端挂载点不可访问
- 原因分析
  - 该问题是由于新加入的文件系统没有mount，导致该节点NFS服务为faulted 状态，从而使客户端无法访问共享
- 处理步骤
  - 首先在support账户使用cfsmntadm display下查看文件系统挂载情况，如果mount，在master账户使用nfs server start命令重启NFS服务即可修复

 Page 46 HUAWEI

- 如果没有mount，在master下使用fs online fs_strip使得文件系统上线，在使用nfs server start命令重启NFS服务修复
- NFS服务恢复正常后，等待约5分钟后（若节点数和网口数较多，请适当增加等待时间），如果IP依然没有分布到其他节点，需要手动使用ip addr online命令重新分布IP

总结

- N8500故障处理的原则、方法
- N8500故障处理的流程
- N8500分类故障以及典型问题处理思路
- N8500故障处理案例

思考题

- N8500的故障分类主要可以分为哪几类?
- N8500的主要日志有哪些？分别在什么路径下?
- N8500内部网络链路异常诊断时需要注意什么?
- N85000文件系统共享异常通常的原因有哪些?

Thank you

www.huawei.com

HC1209206 大数据存储系统维护与故障诊断

HC1209206
大数据存储系统维护

www.huawei.com

HUAWEI

目标

- 学完本课程后，您将能够:
  - 掌握OceanStor 9000系统日常维护方法：巡检、信息收、状态检查

目 录

安装Toolkit

步骤1：获取工具安装包，解压HUAWEI OceanStor Toolkit V100R001CxxSPCxxx.zip包。

说明： 解压到的路径不能包含中文字符。

步骤2：双击解压文件中的start.exe打开工具。

HUAWEI

OceanStor Toolkit是OceanStor 9000的软件工具。

巡检1

步骤1：单击“设备”→“添加”，输入存储系统的管理IP地址，然后单击“下一步”。



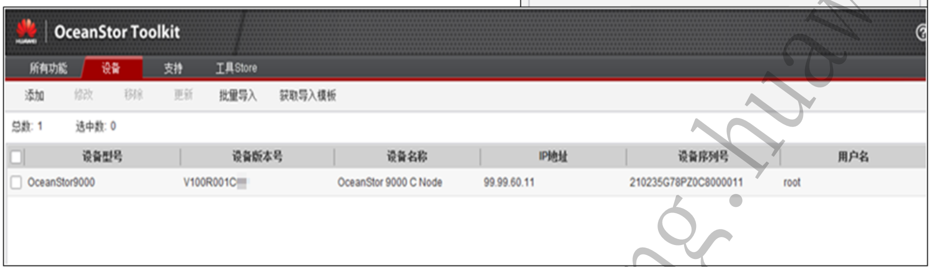
HUAWEI

巡检2

步骤2：在“用户名”中输入“root”，在“密码”中输入“Root@storage”后单击“完成”。

步骤3：单击“完成”，显示设备列表。





HUAWEI

巡检3

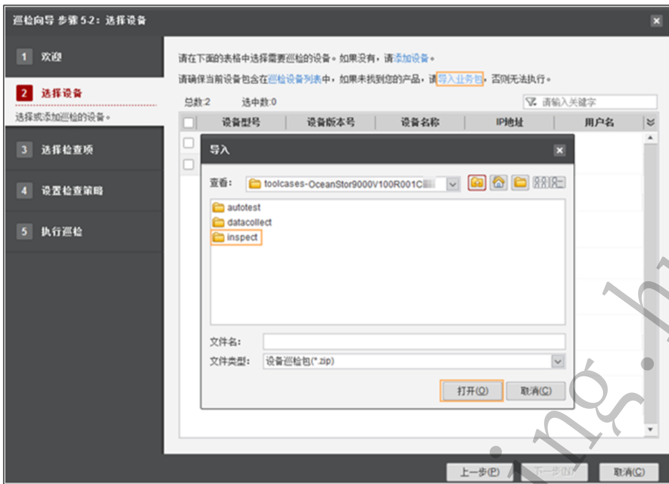
步骤4：选择“所有功能 > 维护 > 巡检及续保检查”。

步骤5：选择“实时巡检”，单击“下一步”。



巡检4

步骤6：单击“导入业务包”，选择当前版本配套的巡检业务包。



 Page 7 HUAWEI

- 说明：业务包在toolcases-OceanStor9000V100R001Cxx.zip的“inspect”目录中。

巡检5

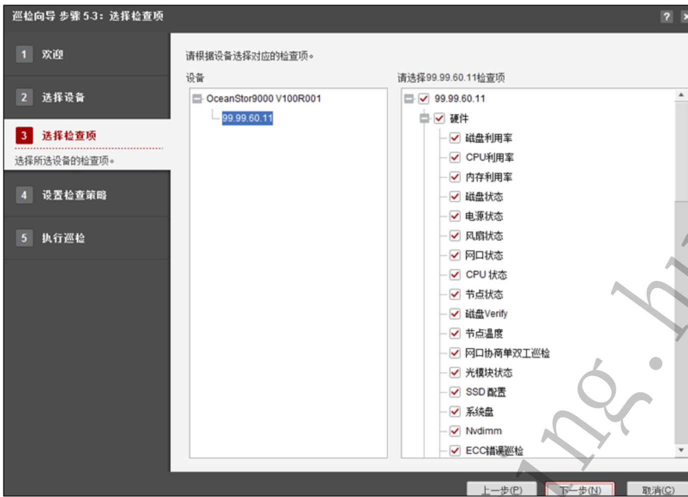
步骤7：选择需要巡检的设备后，单击“下一步”。



 Page 8 HUAWEI

巡检6

步骤8：界面列出了所有的巡检项，根据需求进行选择，默认全部进行巡检。



HUAWEI

巡检7

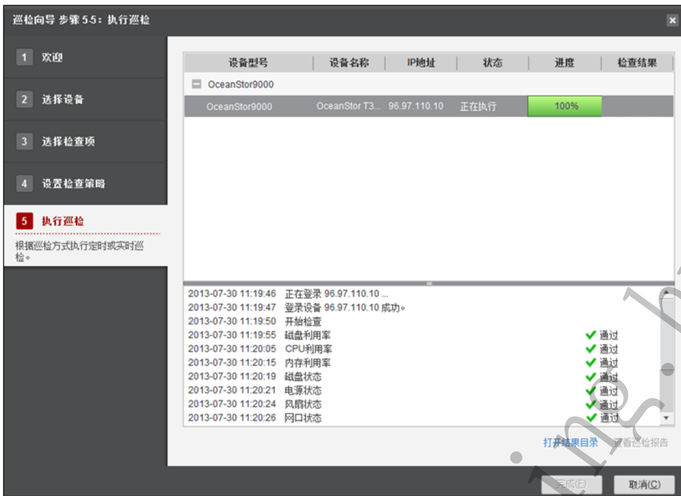
步骤9：选择巡检结果保存路径，单击“下一步”。



  HUAWEI

巡检8

步骤10：待巡检完成后，单击“查看巡检报告”可以进行查看。



目 录

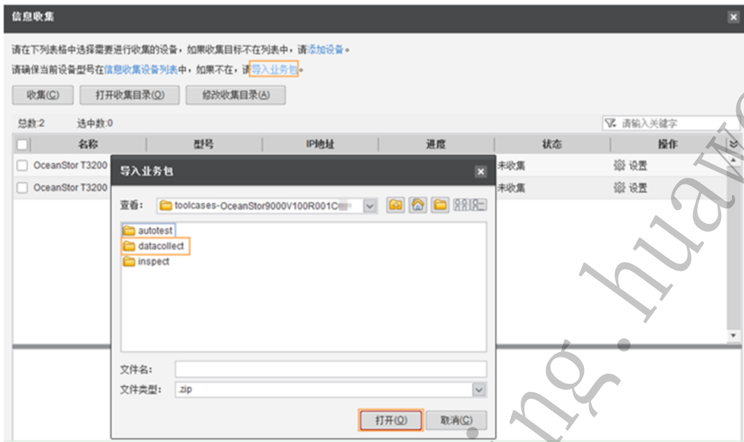
信息收集1

步骤1：选择“所有功能->维护->信息收集”。



信息收集2

步骤2：单击“导入业务包”，选择当前版本配套的信息收集业务包。



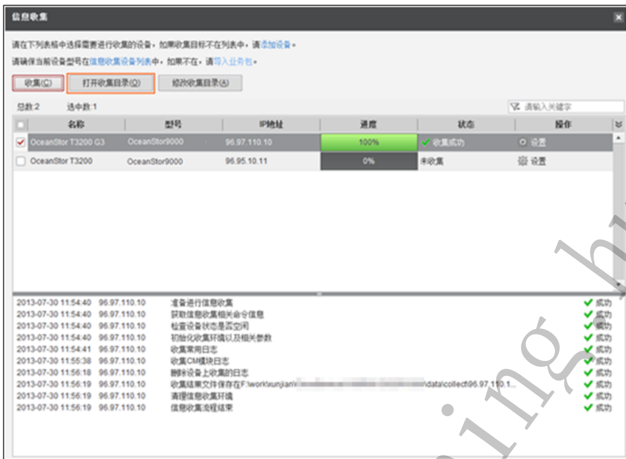
HUAWEI

- 说明：业务包在toolcases-OceanStor9000V100R001Cxx.zip的“datacollect”目录中。

信息收集3

步骤3：勾选待收集的设备后单击“收集”。

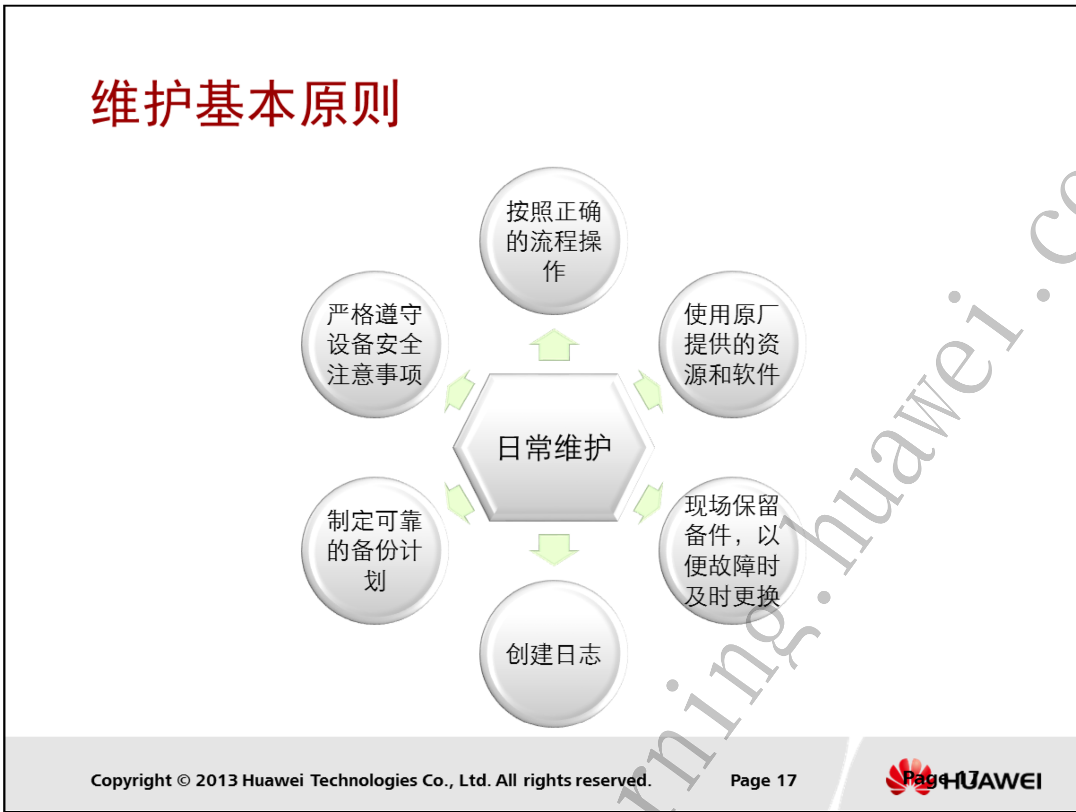
步骤4：当进度变成100%时，单击“打开收集目录”查看信息收集结果。



HUAWEI

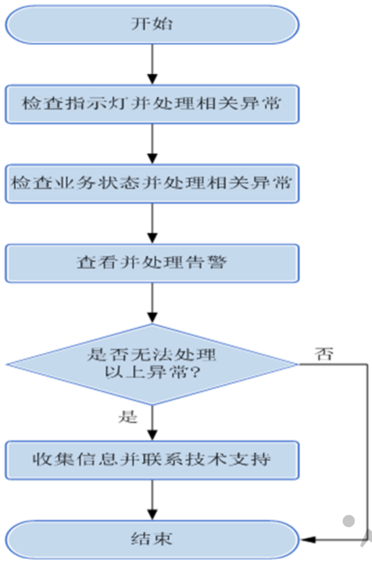
目 录

  HUAWEI



- 通过例行维护能够确认设备环境、设备状态是否正常，并对异常状态进行及时处理维护以保证设备的正常运行。
  - 通过检查设备的当前状态，实时掌握设备和网络的运行状况，了解设备或网络的运行趋势，提高维护人员对突发事件的处理效率。
  - 及时发现设备所发出的告警或已存在的缺陷，并采取适当的措施予以恢复和处理，维持设备的健康水平，降低设备的故障率。

维护流程



开始
检查指示灯并处理相关异常
检查业务状态并处理相关异常
查看并处理告警
是否无法处理以上异常?
否
是
收集信息并联系技术支持
结束

HUAWEI

维护项目

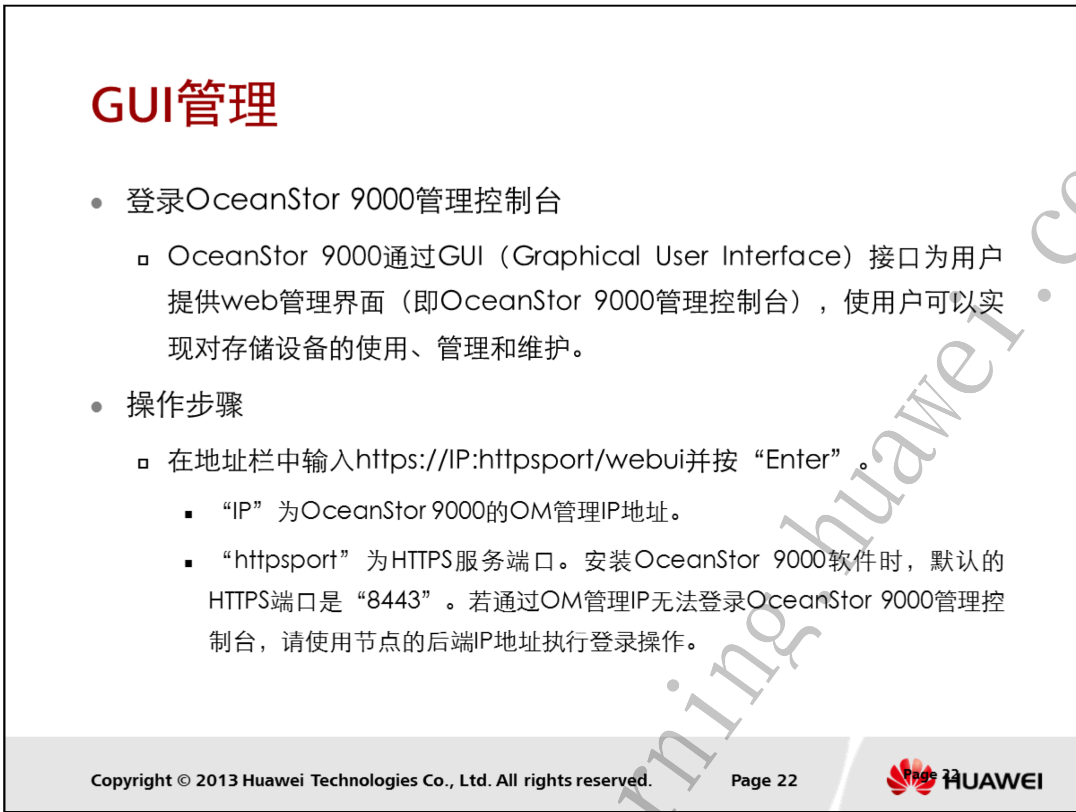
维护地点	维护项目
设备现场侧	检查机房环境，检查机柜内部环境 检查设备的指示灯状态
IPMI设备管理软件侧	检查设备，检查硬盘，检查传感器列表 检查用户信息
	检查告警信息
OceanStor 9000管理控制台侧	查看集群，查看节点，查看节点文件系统服务状态 查看CIFS共享服务状态，查看NFS共享服务状态 查看交换机状态 查看系统操作日志信息 查看性能信息，查看告警、事件信息

维护工具

工具名称	用途
设备管理软件	查询设备运行状况。
OceanStor 9000管理控制台	查询系统、业务运行情况。
光纤清洁工具	使用专用的清洁工具和材料，以免损坏光纤接头。
温度计	测量机房的环境温度。
湿度计	测量机房的环境湿度。

管理方式

- GUI（Graphical User Interface）
  - 服务器端部署在OceanStor 9000上，用户通过终端浏览器连接到OceanStor 9000所在业务网，访问OceanStor 9000进行管理和维护。管理控制台是存储单元管理平台，支持同时管理集群和节点。
- CLI（Command Line Interface）
  - CLI是命令行接口，能够高效、快捷地管理和维护设备。



- 前提条件
  - OceanStor 9000管理控制台所在节点正常运行。
  - OceanStor 9000管理控制台所在节点与其他节点之间通信正常。
  - 维护终端的操作系统与浏览器满足兼容性要求。
  - 浏览器支持Internet Explorer（IE8.0或IE9.0版本）、Safari5、FireFox11和Chrome14或更高版本。

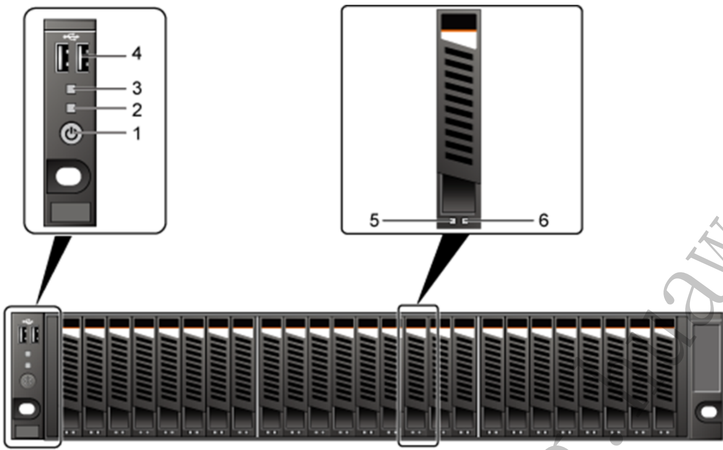
CLI管理

- CLI管理
  - OceanStor 9000通过CLI（Command Line Interface）接口为用户提供管理模块，使用户能够实现对存储设备的使用、管理和维护。
- 登录CLI界面
  - 在Windows操作系统下，用户可以使用SSH客户端软件登录CLI；
  - 在Linux操作系统下，用户可以在Shell中使用“ssh ” “任意节点业务网口的IP地址”命令，登录CLI。

 Page 23 HUAWEI

- 操作步骤
  - 运行PuTTY，在“Host Name(or IP address)”中输入OceanStor 9000的任意节点业务网口的IP地址，其他参数使用默认值，单击“Open”
  - 依次输入用户名“admin”和密码“Admin@123”，按“Enter”后进入操作系统命令行。Admin@123”为操作系统的默认密码，用户登录后可自行修改。
  - 输入cd /startup_disk/image/ISM/ism_ap/CLI/ismcli/命令，按“Enter”。
  - 输入./ismcli -u admin -p Admin@123 -ip 127.0.0.1 -port 8080命令，按“Enter”，登录CLI。
  - 当命令提示符变为“cli:/>”，即可执行CLI命令。
- 前提条件
  - 已使用网线将维护终端连接到OceanStor 9000的交换机。
  - 已设置维护终端的IP地址与OceanStor 9000的业务网络在同一网段。
  - 若维护终端使用Windows操作系统，已获取SSH（Secure Shell）客户端软件。
  - 维护终端在Windows操作系统使用SSH客户端软件PuTTY0.60为例，说明登录CLI的步骤。

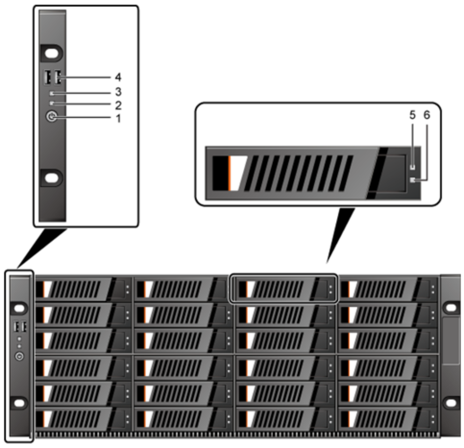
检查P节点/M节点设备运行指标灯



1	系统电源指示灯/开机按钮	2	系统告警指示灯
3	系统定位指示灯	4	USB端口
5	硬盘运行指示灯	6	硬盘告警/定位指示灯

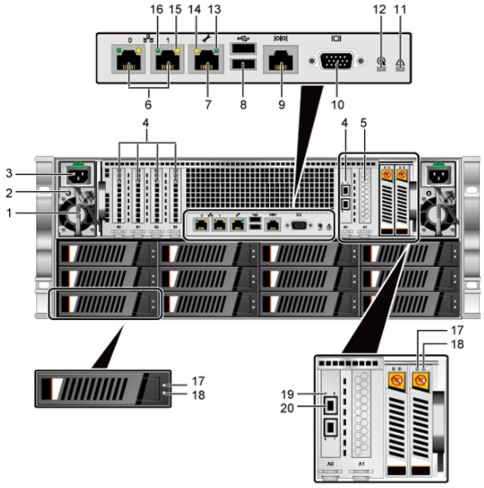
HUAWEI

检查C节点设备运行指标灯



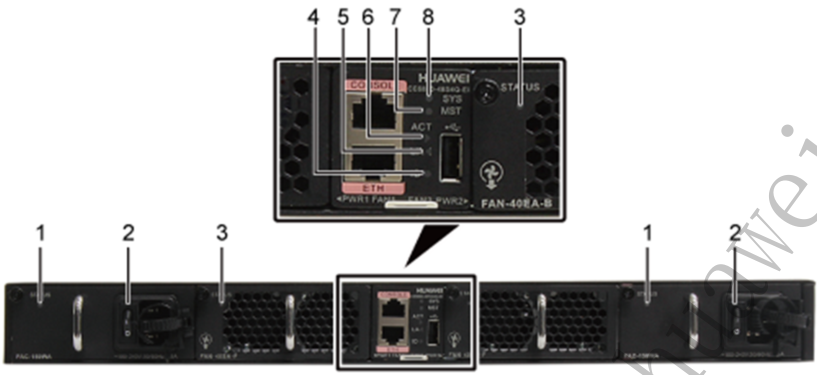
1	系统电源指示灯/开机按钮	2	系统告警指示灯
3	系统定位指示灯	4	USB端口
5	硬盘运行指示灯	6	硬盘告警/定位指示灯

检查C节点设备运行指标灯



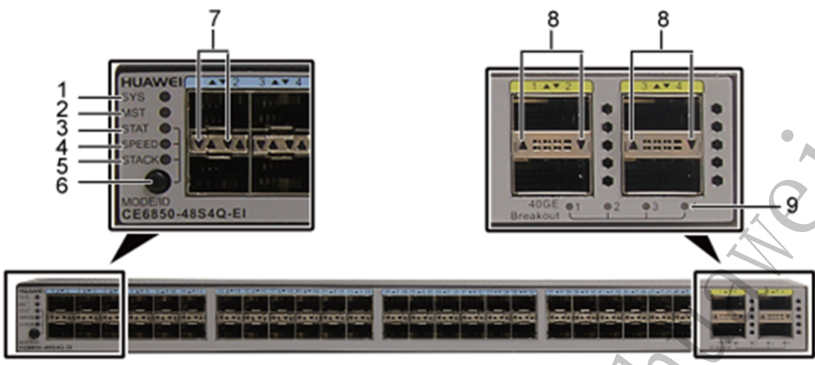
1	电源模块拉手	2	电源运行/告警指示灯
3	电源接口	4	PCI-E扩展插槽
5	SAS卡	6	GE网口
7	IPMI管理网口	8	USB端口
9	系统串口	10	VGA端口
11	系统告警指示灯	12	系统定位指示灯
13	IPMI管理网口LNK指示灯	14	IPMI管理网口ACT指示灯
15	GE网口ACT指示灯	16	GE网口LNK指示灯
17	硬盘运行指示灯	18	硬盘告警/定位指示灯
19	10GE SFP+以太网光接口指示灯	20	10GE SFP+以太网光接口

检查交换机指标灯



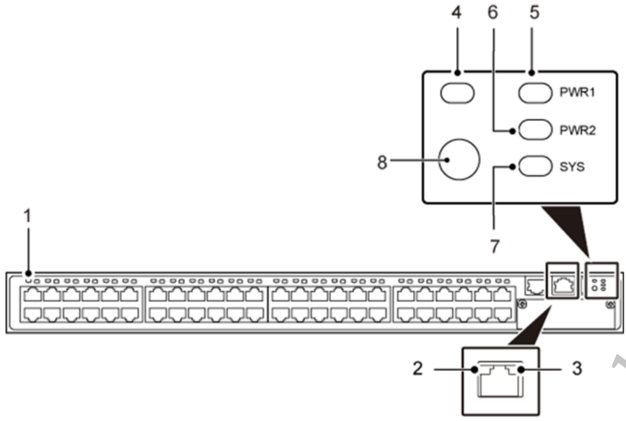
1	电源运行指示灯	2	电源开关
3	风扇运行指示灯	4	MODE模式切换按钮/ID指示灯
5	管理网口指示灯	6	USB开局指示灯
7	堆叠主从指示灯	8	系统运行状态灯

检查交换机指标灯



1	系统运行状态灯	2	堆叠主从指示灯
3	STAT模式状态灯	4	SPEED模式状态灯
5	STACK模式状态灯	6	模式切换按钮/ID指示灯
7	业务接口指示灯（10GE光口）	8	业务接口指示灯（40GE光口）
9	40GE接口的10GE模式指示灯（1/2/3/4）	-	-

检查交换机指标灯



1	以太网电接口工作状态指示灯	2	数据传输状态指示灯
3	连接状态指示灯	4	模式切换指示灯（MODE）
5	电源指示灯（PWR1）	6	电源指示灯（PWR2）
7	系统运行状态指示灯（SYS）	8	模式切换按钮

查看集群状态

- **查看集群状态**
  - 介绍通过OceanStor 9000管理控制台检查集群状态的具体操作。通过检查集群的状态，可以及时发现和处理故障。
- **参考标准**

  在OceanStor 9000管理控制台中:
  - 集群的运行状态为“运行”。
  - “当前告警”列表中没有新增与集群相关故障信息。

- 操作步骤

  步骤一 登录OceanStor 9000管理控制台。

  步骤二 在左上方基本信息中查看集群状态。

查看节点状态

- 查看节点状态
  - 介绍通过OceanStor 9000管理控制台检查节点状态的具体操作。通过检查节点的状态，可以及时发现和处理故障
- 参考标准

  在OceanStor 9000管理控制台中:
  - 节点的运行状态为“运行”。
  - “当前告警”列表中没有新增与节点相关故障信息。

 Page 31 HUAWEI

- 操作步骤

  步骤一 登录OceanStor 9000管理控制台。

  步骤二 进入“节点列表”界面。

  在右侧导航中单击 “系统”。

  在“基本信息”区域，单击“查看详情”

  步骤三 在中间信息展示区浏览节点信息。

查看节点文件系统服务状态

- 查看节点文件系统服务状态
  - 介绍通过OceanStor 9000管理控制台检查节点文件系统服务状态的具体操作。通过检查节点文件系统服务的状态，可以及时发现和处理故障。
- 参考标准

  在OceanStor 9000管理控制台中:
  - 节点文件系统服务状态为“正常”。
  - “当前告警”列表中没有新增与节点文件系统服务相关故障信息。

 Page 32 HUAWEI

- 操作步骤

  步骤一 登录OceanStor 9000管理控制台。

  步骤二 进入“节点列表”界面。

  在右侧导航中单击 “系统”。

  在“基本信息”区域，单击“查看详情”

  步骤三 单击某个节点的名称进入节点基本信息界面。

  步骤四 在节点基本信息中查询此节点“文件系统服务运行状态”是否为“正常”。

总结

- OceanStor 9000存储系统巡检、信息收集和日常维护方法与工具

HUAWEI

思考题

- OceanStor 9000系统巡检和信息收集工具和方法
- OceanStor日常维护状态检查，需要重点关注那些项目

习题

练习题：

1、（判断题）Tookkit维护工具除了可以进行设备巡检以外，还可以进行信息收集。

2、（选择题）9000存储系统日常维护，主要检查哪些项目：（　　）

A、硬盘状态

B、节点和交换机状态

A、集群状态

B、文件系统共享服务

- 1、T；2、ABCD

Thank you

www.huawei.com

HC1209207 集群 NAS 存储系统规划与设计

HUAWEI

HC1209207
集群NAS存储划与设计

www.huawei.com

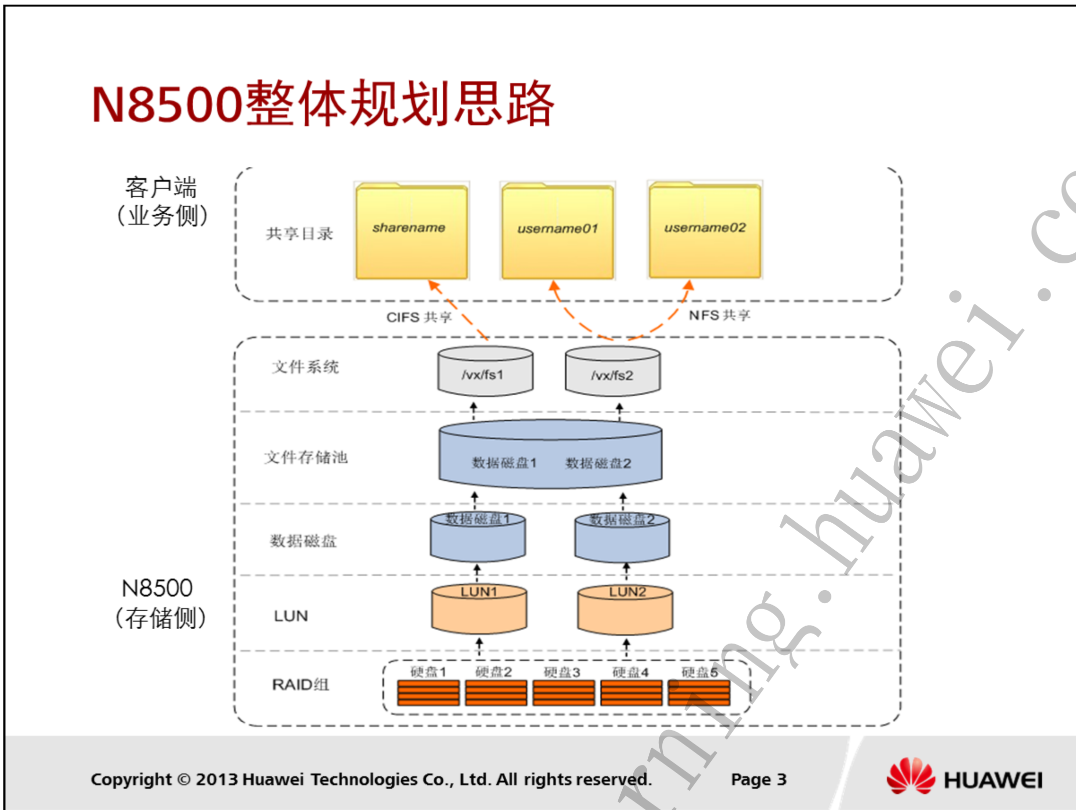
HUAWEI

目标

- 学完本课程后，您将能够:
  - 描述N8500整体规划思路
  - 掌握N8500网络规划
  - 掌握N8500容量规划
  - 掌握N8500文件系统规划
  - 掌握N8500业务规划
  - 了解N8500性能优化

HUAWEI

目录



N8000存储层级自下而上分别为RAID、LUN、数据磁盘、文件存储池和文件系统。规划时需考虑以下几点：

- 业务侧：负载尽量均衡到集群各节点，多网口绑定增强网络路径带宽及保护性。
- 存储侧：文件系统为NAS共享业务的基本存储单元。客户端以共享的方式使用文件系统。客户环境对性能、数据安全性的要求，需规划不同类型的文件系统。
- 存储侧： LUN由RAID组创建，扫描后识别为数据磁盘，是用于创建文件系统的基本存储单元。
- 存储侧： RAID组由若干硬盘组成，不同级别的RAID组可以满足不同应用场景对数据安全性或读写性能的要求。

文件存储池是数据磁盘的集合，主要为创建文件系统提供统一的存储空间。创建文件存储池便于整合和分配存储空间。

目录

HUAWEI

N8500网络IP规划

规划项	IP地址信息/子网掩码	网关
引擎节点物理IP地址段	举例：192.168.10.10/24～192.168.10.10/24	192.168.10.1
引擎节点虚拟IP地址段	________～________	
引擎节点Console IP地址	________	
控制框管理网口IP地址	控制器A：________控制器B：________	________

规划项	数量计算方法	网段要求
引擎节点物理IP	组建业务网络的网口（考虑扩展）数量+ 引擎节点所有管理网口数量。	•必须为连续的IP地址段。 •所有IP地址（Console IP地址、物理IP地址、业务网口虚拟IP地址）在同一网段内。
引擎节点虚拟IP	引擎节点业务网口的数量。	•必须为连续的IP地址段。 •所有IP地址（Console IP地址、物理IP地址、业务网口虚拟IP地址）在同一网段内。
引擎 Console IP	每套N8500只需规划1个。	•与控制框的控制器IP地址在同一网段。 •与引擎物理IP地址、业务网口虚拟IP地址在同一网段内。
控制框管理IP	多台控制框的情况下，需对每一台控制框规划不同的管理网口IP地址。	与引擎节点Console IP地址在同一网段内。

HUAWEI

N8000的地址分两类，一类为物理IP地址，一类为虚拟IP地址。

- 物理IP地址在所有业务网口和引擎节点管理网口、存储单元管理网口各有一个；
- 虚拟IP地址分为业务网口虚拟IP地址（每个业务网口一个）和管理IP地址（Console IP，所有引擎节点管理网口共用一个）。

在做IP地址规划时，需要保证每个端口上的物理IP和虚拟IP在同一个网段。如果您初始化完成后想调整管理网口IP地址与业务网口IP地址不在同一段，请同时调整虚拟IP和物理IP。

部署时，建议N8000与客户现网不连通（如果使用临时IP地址部署然后调整的方案，请务必先保证不与客户网络连通），避免出现相互影响，待部署完成后需要调测业务时再连通。

- 对于IP资源，切换规则如下：
    1. 从一个引擎节点切换到另外一个引擎节点，不能在单个引擎节点的多个网口之间切换。
    2. 在引擎节点间的相同网口名的网口间切换，不能从一个引擎节点的网口1切换到另外一个引擎节点的网口2
- 根据ip切换规则，对组网有如下要求：
    1. 集群中各个引擎节点相同网口名的网口需连接到同一网段中（比如所有节点的pubeth1端口必须同一网段）。
    2. 同一个引擎节点的不同网口可以连接到不同的网段，也可以连接到同一网段中

注意：表中规划的IP不包括阵列提供SAN服务时的ISCSI端口的IP。

N8500网络IP规划

规划项	IP地址信息/子网掩码	网关
引擎节点物理IP地址段	举例：192.168.10.10/24～192.168.10.10/24	192.168.10.1
引擎节点虚拟IP地址段	________～________	
引擎节点Console IP地址	________	
控制框管理网口IP地址	控制器A：________ 控制器B：________	________

规划项	数量计算方法	网段要求
引擎节点物理IP	组建业务网络的网口（考虑扩展）数量+ 引擎节点所有管理网口数量。	•必须为连续的IP地址段。 •所有IP地址（Console IP地址、物理IP地址、业务网口虚拟IP地址）在同一网段内。
引擎节点虚拟IP	引擎节点业务网口的数量。	•必须为连续的IP地址段。 •所有IP地址（Console IP地址、物理IP地址、业务网口虚拟IP地址）在同一网段内。
引擎 Console IP	每套N8500只需规划1个。	•与控制框的控制器IP地址在同一网段。 •与引擎物理IP地址、业务网口虚拟IP地址在同一网段内。
控制框管理IP	多台控制框的情况下，需对每一台控制框规划不同的管理网口IP地址。	与引擎节点Console IP地址在同一网段内。

 Page 6 HUAWEI

如客户无特殊要求，建议配置管理网口Console IP地址和业务网同一个网段，即所有物理IP、虚拟IP、管理IP一个段内，方便维护。

- 规划时须注意：
  - 部署前使用串口以support用户登录设备，执行cat /etc/HS/conf/version.conf检查设备版本，如果低于N8000V200R001C00SPC104，请及时联系研发工程师升级系统版本。（默认密码support@storage）
  - 用户现网网段不满足要求时需先用满足网段要求的临时方案初始化，然后修改地址为实际规划
  - ISM ServiceTool部署工具适用于所有的虚拟IP地址、物理IP地址、管理地址在同一网段的情况，如果不满足，请采用将所有IP都设置为管理网段IP的临时方案部署，部署完成后，使用ISM或network模式下IP addr modify修改业务网段地址为规划的正确地址。

N8500网络IP规划举例

规划项	IP地址信息/子网掩码	网关
引擎节点 物理IP地址段	192.168.10.2 /24 ～ 192.168.10.81 /24	192.168.10.1
引擎节点虚拟IP地址段	192.168.10.82 /24～ 192.168.10.161 /24	192.168.10.1
引擎节点Console IP地址	192.168.10.2 /24	192.168.10.1
控制框管理网口IP地址	控制器A：192.168.10.3 /24 控制器B：192.168.10.4 /24	192.168.10.1

 Page 7 HUAWEI

- 具体计算如下：
  - 引擎节点所需全部物理IP地址数量= 48（1 Gb iSCSI网口数量） + 8（10 Gb iSCSI网口数量） + 4（管理网口数量） + 20（计划扩展数量） = 80个。
  - 1 Gb iSCSI网口数量= 4（引擎节点数量） * 3（1Gb iSCSI接口模块数量） * 4（1 Gb iSCSI接口模块网口数量） = 48。
  - 10 Gb iSCSI网口数量= 4（引擎节点数量） * 1（10Gb iSCSI接口模块数量） * 2（1 0Gb iSCSI接口模块网口数量） = 8。
  - 管理网口数量= 4（引擎节点数量） * 1（管理网口数量）。
  - 计划扩展数量根据客户实际情况确定，此处以预估为20个为例。

N8500配置4个引擎节点和1个控制框。除去组建心跳网络的1Gb iSCSI接口模块（不需配置IP地址），每个引擎节点配置3个1Gb iSCSI接口模块（每个接口模块4个网口），1个10Gb iSCSI接口模块（每个接口模块2个网口），用于组建业务网络；每个控制框配置2个控制器。网络IP地址规划如上表。

N8000配置——bond4模式

- 什么是bond4?
  - 即“802.3ad”模式，动态链接聚合策略。
- 优点
  - 网卡资源利用率高
  - 实现负载平衡
  - 提供容错功能
  - 实现数据包发送的有序传输
- 缺点
  - “网关型”拓扑下，负载不均衡问题突出
  - 使用范围有限制

 Page 8 HUAWEI

- 优点
  - 网卡资源利用率高：由于设备链表中的每一块网卡都有机会发送数据包，因此bonding设备中网卡资源的利用率是100%。
  - 实现负载平衡：在“802.3ad”模式下，多个网卡端口工作在同一个激活的聚合体下，共享相同的速率和双工模式。避免了一块网卡长期发送数据，而导致其它网卡失效。
  - 提供容错功能：当bonding设备驱动检测到其中一块网卡出错时，该网卡会被bonding驱动从slave设备链表中清除。它的传输任务将自动由链表中下一块网卡接管，因此不会影响数据发送。
  - 实现数据包发送的有序传输：802.3ad标准要求帧按顺序（一定程度上）传递，因此单个连接一般不会看到包的乱序
- 缺点
  - 在“网关型”拓扑下，由于“802.3ad”模式的实现通过对端来分发流量（通过MAC 地址的XOR 值），所有外出流量将使用同一个网卡设备，负载不均衡问题突出。
  - “802.3ad”模式要求所有设备在同样的速率和双工模式下进行聚合，限制了该模式的使用范围。

N8000配置——bond6模式

- 什么bond6?
  - 即“balance-alb”模式，自适应负载均衡策略。
- 优点
  - 能够根据bonding设备的负载状态进行动态的分配和调整。
  - 提供容错功能。
- 缺点
  - 动态平衡必然会造成网卡设备的频繁切换，因而增加了系统开销。
  - ARP 协商机制的更新(服务器端)的方式不够健壮，有可能出现bonding设备主网卡承担过多负载的情况。

Page 9
HUAWEI

在提供传输负载均衡功能的基础上，“balance-alb”模式增加针对IPV4 流量的接收负载均衡（receive load balance， rlb）。Bonding驱动程序根据每一个bonding网卡端口（slave）的不同处理能力和当前的负载状态，对网络服务请求进行合理的调配，将网络发送和传输的负载在所有网卡端口中重新分布。

- 优点
  - 能够根据bonding设备的负载状态进行动态的分配和调整。“balance-alb”模式根据每一个平衡周期内网络网卡的负载不同，在网卡组内动态分配发送和接收任务，避免了某些网卡端口承载大量发送或接收请求，从而在一定程度上提高了服务器和客户端之间的整体吞吐率。
  - 提供容错功能。当bonding设备驱动检测到其中一块网卡出错时，该网卡会被bonding驱动从slave设备链表中清除。它的传输任务将自动由链表中下一块网卡接管，因此不会影响数据发送。
- 缺点
  - 动态平衡必然会造成网卡设备的频繁切换，因而增加了系统开销。
  - ARP 协商机制的更新(服务器端)的方式不够健壮，有可能出现bonding设备主网卡承担过多负载的情况。

N8000配置——配置bond

- 创建bond

```
CFSN8K> network bond create pubeth1,pubeth2 6
100% [#] Bonding interfaces. Please wait...
bond created, the bond name is: bond0
```

- 删除bond

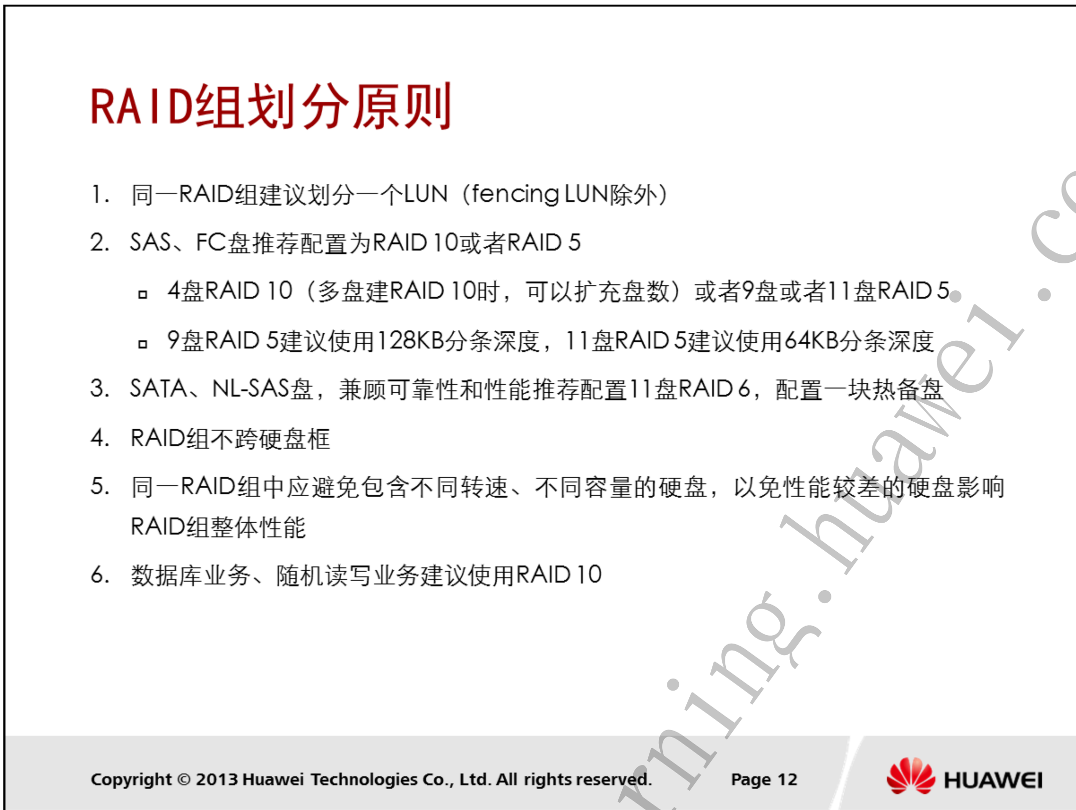
```
CFSN8K> network bond remove bond0
100% [#] Removing Bond bond0. Please wait...
bond removed : bond0
```


HUAWEI

注意：不允许在在线业务上使用bond命令，会引起业务中断。

目录

HUAWEI



- Raid的选型和配置建议：
  - FC/SAS盘建议配置Raid5
  - NL_SAS/SATA必须配置Raid6
- N8000有效容量计算方法：
  - 精确计算方式：有效容量=Raid组个数*（单个Raid组盘数-1(Raid5)或2(Raid6)）*单盘容量*0.93*0.99

  注：0.93为硬盘格式化后可用比例，0.99是创建文件系统可用比例
  - 粗略计算方式：有效容量≈裸容量×0.7

存储划分原则——LUN参数详解

参数项	参数	参数说明	应用场景
预取策略	不预取	按照I/O请求中指定的读取长度进行读取，即不预取。	随机读业务，建议采用“不预取”，否则可能由于读命中率低而会导致性能下降。
	智能预取	智能预取会对主机读请求进行连续性判断，如果是连续的请求，则将当前读请求后面的一段数据从硬盘预取到Cache中，提高读Cache命中率。如果是随机读，则不预取，只从硬盘读取需要的部分。	适用于顺序读和随机都存在，或者无法判断读应用是否顺序的情况。
	固定预取	Cache读取硬盘数据时，每次从硬盘中读取固定的长度（长度由用户配置）。	适合大小固定的顺序读应用。例如彩铃、多用户流媒体点播，用户使用相同码流。
	可变预取	在读取数据时，按照主机I/O请求中的读取长度的倍数将数据预取到Cache中。	可适合大小不固定的顺序读应用，或多用户并发读但无法 判定预读量，无法按固定预取设置的应用。例如多用户多媒体点播， 用户使用码流各不相同。

HUAWEI

存储划分原则——LUN参数详解

参数项	参数	参数说明	应用场景
写策略	透写	数据直接写入硬盘。每次写操作都要访问硬盘。	1. 可靠性最高，但写性能较差，不影响读性能。 2. 适用于对写性能要求不高，对可靠性要求较高的应用。
	回写	1. 回写/镜像：数据写入本地Cache，同时写到对端Cache。 2. 回写/不镜像： 数据只写入本地Cache。	1. 回写、镜像：写性能比透写好，但比回写不镜像差，不影响读性能。适用于写性能和可靠性兼顾的应用。 2. 回写、不镜像：写性能最好，但可靠性较差。适用于对写性能要求高，对可靠性要求低的应用。
	强制回写	1. 强制回写、镜像：当存储系统发生某些故障（例如高温故障或BBU供电不足）时，强制写入本地Cache，同时写到对端Cache。 2. 强制回写、不镜像：当存储系统发生某些故障（例如高温故障或BBU供电不足）时，强制写入本地Cache。	1. 强制回写、镜像：写性能比透写好，但比强制回写不镜像差，不影响读性能。系统故障时可能会有数据丢失的风险。 2. 强制回写、不镜像：写性能最好，但可靠性较差，不影响读性能。系统故障时可能会有数据丢失的风险。

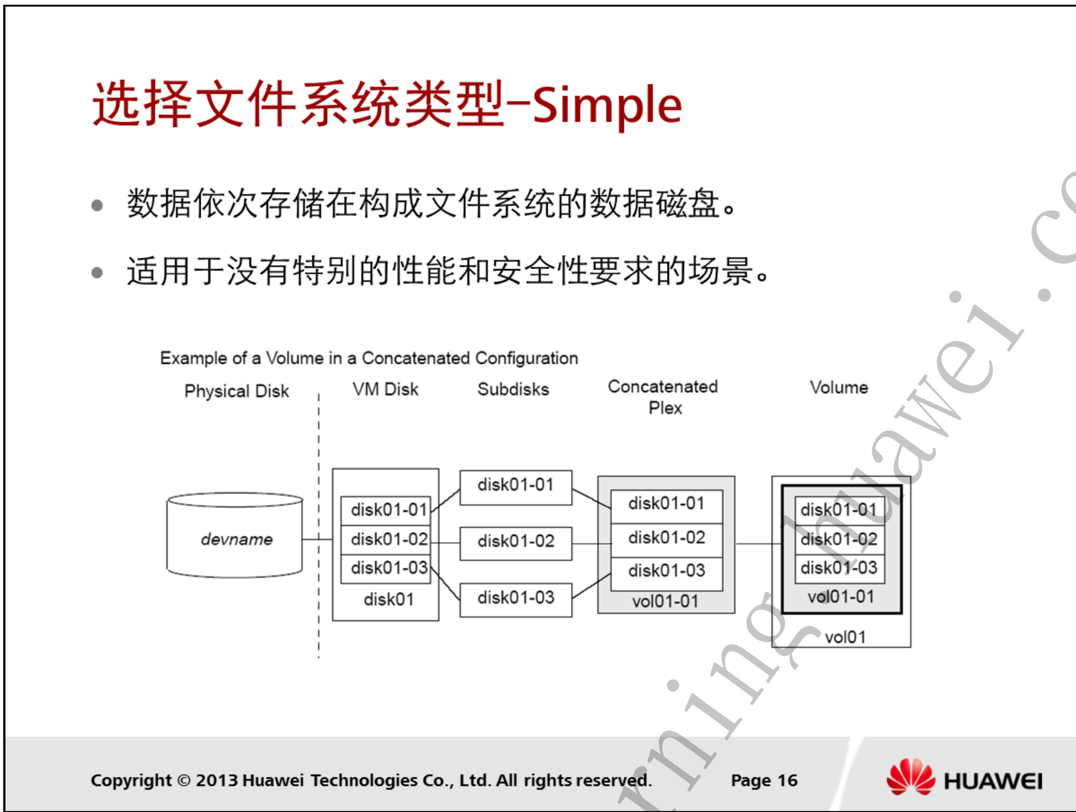
 Page 14 HUAWEI

- **说明：**当写策略设置为“回写”或“强制回写”时，才能与镜像策略搭配使用。

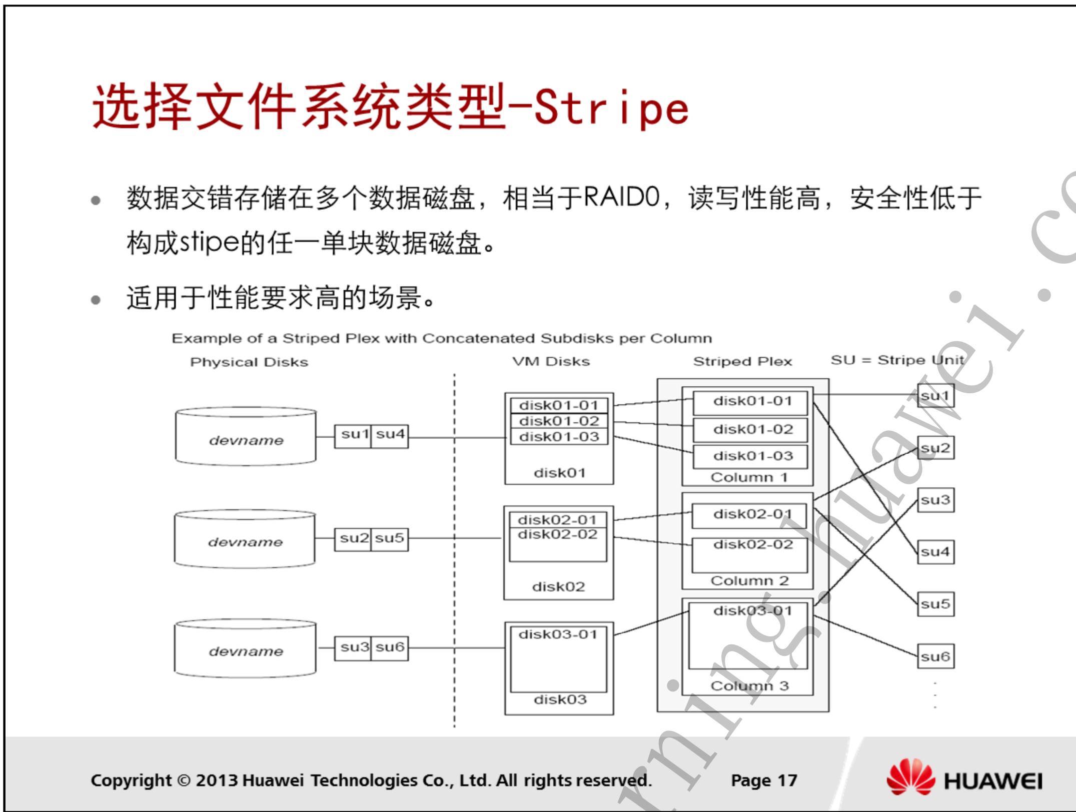
存储划分原则——LUN

1. 单个LUN推荐最大不超过16TB
2. LUN归属控制配置：同一RAID组的LUN归属于同一控制器，不同的RAID组交替归属于A控和B控。
3. 写cache：采用回写镜像
4. 预取策略请参考“LUN参数详解”
5. LUN映射：映射时主机LUN ID不能为0
6. LUN分条深度最小值不建议小于32K，顺序型业务建议使用64K分条深度，随机型业务、和混合型业务建议使用32K分条深度
7. 如果有多个业务量较大的LUN，建议规划在不同的RAID组上。

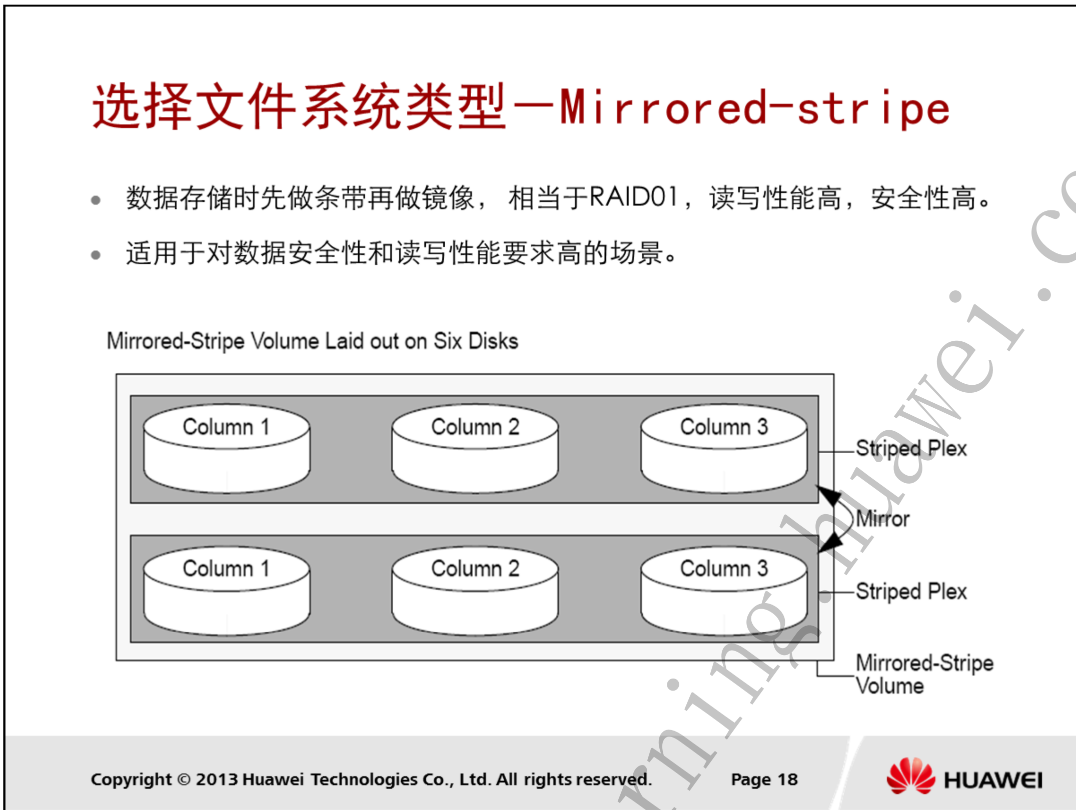
  HUAWEI



Simple文件系统的数据依次存储在构成文件系统的数据磁盘（文件系统的卷可能有来自多个数据磁盘的子磁盘构成），首先将从头到尾访问第一个subdisk上的数据，然后在从头到尾访问下一个subdisk上的数据，直到访问完毕所有的subdisk。相当于单块数据磁盘，读写性能及容错性与单块数据磁盘一致。



条带化数据的方法可以使数据在两个或多个磁盘上交错分布。条带化plex包括两个或多个subdisk，这些subdisk来自两个或多个磁盘。数据会在这些subdisk上交错分配。这些subdisk组织成列，每列都包含来自一个或多个物理磁盘的subdisk。每列的subdisk的数量和大小可能不同。在必要时也可向列中添加额外的subdisk。例如在条纹化的plex中有三列和六个条纹单元，数据在这三列上的条纹化分布将如图所示。



VxVM支持在条带化的基础上施加映像。图例显示了两个plex，每个都有跨越3个磁盘的条带，它们在一起互为映像，组成了一个Mirrored-stripe卷。

文件系统参数规划

1. 对于条带类型的文件系统，columns 数量不超过8，防止数据过于离散。
2. 文件系统大小超过1T时blocksize选择8192。
3. 创建条带文件系统，建议columns数量与raid组个数一致。
4. 在实际配置时，从可靠性和故障恢复速度方面考虑，文件系统大小建议不超过64TB。具体视文件大小决定，对小文件（平均文件大小<1M）存储应用，不超过16TB，文件越小，在满足容量需求的情况下，建议创建越小容量的文件系统。
5. 文件数量超过1000万时，建议不开启配额。
6. 建议单个文件系统存放文件不超过4000万，目录层度不超过10层.
7. 建议每个目录下不超过1万个文件，不超过1千个目录。

目录

文件共享业务规划原则

1. NFS共享支持集群式访问，主机挂载需要手动分配IP，为负载均衡，请使用浮动在不同节点的虚拟IP。
2. CIFS Normal共享采用的是DFS，单一文件系统只能通过特定机头访问，推荐ctdb集群模式。
3. 10GE端口不建议绑定，1GE端口绑定建议模式选择模式4或模式6。

虚拟IP均衡分布

- Network->ip addr show 命令查看IP分布

```
CFSN8K> network ip addr show
IP           Netmask       Device    Node          Type     Status
--           -------       ------    ----          ----     ------
129.88.51.60  255.255.0.0  pubeth0  CFSN8K_01   Physical
129.88.51.61  255.255.0.0  pubeth1  CFSN8K_01   Physical
129.88.51.62  255.255.0.0  pubeth2  CFSN8K_01   Physical
129.88.51.63  255.255.0.0  pubeth3  CFSN8K_01   Physical
129.88.51.64  255.255.0.0  pubeth4  CFSN8K_01   Physical
129.88.51.65  255.255.0.0  pubeth0  CFSN8K_02   Physical
129.88.51.66  255.255.0.0  pubeth1  CFSN8K_02   Physical
129.88.51.67  255.255.0.0  pubeth2  CFSN8K_02   Physical
129.88.51.68  255.255.0.0  pubeth3  CFSN8K_02   Physical
129.88.51.69  255.255.0.0  pubeth4  CFSN8K_02   Physical
129.88.51.51  255.255.0.0  pubeth0  CFSN8K_02   Virtual  ONLINE (Con IP)
129.88.51.80  255.255.0.0  pubeth1              Virtual  OFFLINE
129.88.51.81  255.255.0.0  pubeth1              Virtual  OFFLINE
129.88.51.82  255.255.0.0  pubeth2              Virtual  OFFLINE
129.88.51.83  255.255.0.0  pubeth2              Virtual  OFFLINE
129.88.51.84  255.255.0.0  pubeth3              Virtual  OFFLINE
129.88.51.85  255.255.0.0  pubeth3              Virtual  OFFLINE
129.88.51.86  255.255.0.0  pubeth4              Virtual  OFFLINE
129.88.51.87  255.255.0.0  pubeth4              Virtual  OFFLINE
```

 Page 22 HUAWEI

- 虚拟IP地址在各NAS节点上分布均匀是最佳状态。

NFS共享业务均衡

- 客户端使用不同的IP挂载NFS共享

```
CFSN8K> nfs share show
/vx/fs01                    129.88.* (rw,no_root_squash,async)
CFSN8K> nfs server status
NFS Status  on CFSN8K_01 : ONLINE
NFS Status  on CFSN8K_02 : ONLINE

suse11:~ # mount -t nfs 129.88.51.60:/vx/fs01  /mnt/ip1
suse11:~ # mount -t nfs 129.88.51.65:/vx/fs01  /mnt/ip2
suse11 :~ #mount -t nfs
129.88.51.60:/vx/fs01  on /mnt/ip1  type  nfs (rw,addr=129.88.51.60)
129.88.51.65:/vx/fs01  on /mnt/ip2  type  nfs (rw,addr=129.88.51.65)
```


HUAWEI

CIFS共享业务均衡

- CIFS共享挂载原则
  - 推荐使用集群模式（cdtb），集群模式在同一时刻由多个集群NAS引擎节点为某个CIFS共享提供服务，缓解单节点业务负载，提高访问速度。

```
CFSN8K> cifs server status
CIFS Status on CFSN8K_01 : ONLINE
CIFS Status on CFSN8K_02 : ONLINE

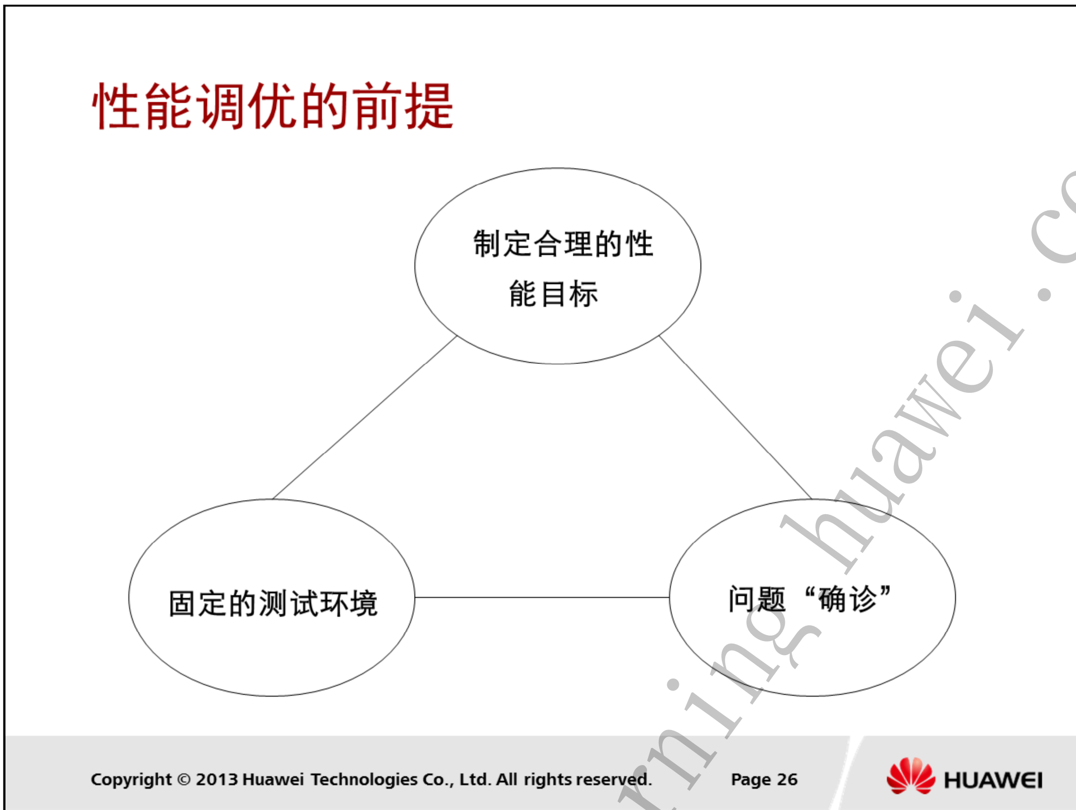
Homedirfs              :
Security               : user
Clustering Mode        : normal

CFSN8K> cifs share show
         ShareName         FileSystem  ShareOptions
              dst               fs_dst   owner=root,group=root,rw
CFSN8K> cifs share show dst
         ShareName         VIP  Address
              dst            129.88.51.80
```


- 为了避免造成性能瓶颈，CIFS共享时需要创建与NAS节点个数相同数量的文件系统

目录

HUAWEI



开始性能调优工作之前，首先要根据现有经验评估性能目标的合理性，当性能目标明显超过了硬件能力的时候，应该考虑增强硬件配置。一般来说，千兆网卡上TCP的最大传输能力约100MB/s，万兆网卡上TCP的最大传输能力约1000MB/s。

在做性能优化调整的时候，要使用一套固定的测试环境，并确定一种固定的测试模式，在这种条件下进行各种试验性的调整，并记录每次调整对性能的影响。如果测试环境和模式无法固定，则测试结果就不具有可比性，测试结果就是无意义的。

在开出“药方”之前，应首先“确诊”。通过不断的“猜测问题原因”、“做实验验证原因”的循环，确定问题根因，然后再给出问题解决方案。

N8500性能评估关键参数

性能评估关键参数	
每个NAS引擎节点提供的网口数量	GE网口数量
	10GE网口数量
NAS引擎节点硬件配置	CPU数量及型号
	内存大小
NAS引擎和存储单元通讯链路	4Gb FC
	8Gb FC
后端存储单元性能	S2600T/S5500T/S5600T/S5800T/S6800T
磁盘类型	SSD/SAS/FC/SATA/NL_SAS

HUAWEI

N8000性能评估依据

性能评估依据	
引擎性能/节点 （大文件混合读写）	N8500基础版（单CPU + 4个千兆网口 + 16GB cache） >350MB/s N8500标准版（单CPU + 万兆网口 + 24GB cache）>500MB/s N8500企业版（双CPU + 万兆网口 + 48GB cache）>700MB/s N8500增强版（双CPU + 万兆网口 + 96GB cache）>900MB/s
存储单元控制器最大性能（发挥存储最大性能需提供足够的磁盘）	S2600T:800MB/s S5500T:2000MB/s S5600T:3000MB/s S5800T:4000MB/s S6800T: 5000MB/s
磁盘性能/单盘 （大文件混合读写）	FC/SAS：30MB/s (8+1 Raid5) SATA：12MB/s （10+2 Raid6) NL_SAS：20MB/s (10+2 Raid6)

  HUAWEI

- Note：性能评估模型在顺序大文件读写的场景下预计可达到上述评估值，仅作为评估经验参考，不作为答标依据和性能承诺
- 顺序大文件读写场景通常是指文件分布在几十MB以上，NFS协议。

N8000性能计算方法

- 性能计算方法
  - P1: 后端存储单元性能 = 单盘性能 * 有效磁盘数（除去热备盘后的磁盘总数）
  - P2: NAS引擎性能 = 单个NAS引擎节点的性能 * 引擎节点数 *80%（后端存储单元性能足够的情况下）
- N8000总体性能= Min（P1，P2）
  - P1：存储单元性能不超出存储单元控制器的最大性能上限
  - P2：引擎数量小于4时按此规则预估

- 仅限于评估参考



1. 检查后端存储的整体性能是否比客户要求低。
2. 检查所有BBU是否都已故障，如果都故障会导致lun的“当前写策略”变为透写模式。
3. 通过ISM检查用于业务的LUN的“当前写策略”是否已变为透写模式。
4. 检查磁盘有无慢盘，登录阵列ISM，检查是否有硬盘响应慢的告警。
5. 光纤端口速率误码LUN：在引擎节点上用iostat –xm 1看到，有些磁盘的利用率一直很高，那么很可能是磁盘IO带宽是性能瓶颈，应首先解决这个问题。在ISM上使用系统配置导航栏下误码率管理（先清除后刷新）查看有无误码。
6. 前端业务端口速率最佳，如测8Gb的带宽端口速率不能工作在4Gb模式。

性能调优检查项——光纤连接

- 磁盘阵列到级联框光纤连接
  - 控制框A控侧EXP连接到级联框A控制侧PRI，控制框B控侧EXP连接到级联框B控侧PRI，不能交叉相连。
- 引擎到磁盘阵列光纤连接
  - 当同一引擎节点需连接阵列的单个控制器的多个端口时，尽量连接不同扣卡端口。
  - 确保有单独的路径到达阵列A控和B控。
  - 引擎侧如果存在多条路径到达阵列时，可在光纤交换机上划分ZONE实现逻辑隔离。

HUAWEI

性能调优检查项

- LUN划分和归属
  - 一个RAID组上仅创建一个LUN
  - LUN处于回写镜像状态。
  - 确保LUN归属控制器与工作控制器一致。
- 控制器分配
  - 组成文件系统的磁盘均衡分配在控制器上，比如可选择条带类型文件系统，而不是使用简单类型的文件系统。

 Page 32 HUAWEI

性能调优检查项——CPU利用率

- 监控cpu利用情况

```
ZKY_1:~ # sar -u 3 3
Linux 2.6.5-7.287.3-smp (ZKY_1)          11/24/09

21:48:26        CPU     %user     %nice   %system   %iowait     %idle
21:48:29        all      0.00      0.00      0.25      0.08     99.67
21:48:32        all      0.00      0.00      0.25      0.92     98.83
21:48:35        all      0.08      0.00      0.33      0.75     98.83
Average:        all      0.03      0.00      0.28      0.58     99.11
```

- 监控内存利用情况

```
ZKY_1:~ # sar -r 3 3
Linux 2.6.5-7.287.3-smp (ZKY_1)          11/24/09

21:50:49    kbmemfree kbmemused  %memused kbbuffers  kbcached kbswpfree kbswpused  %swpused  kbswpcad
21:50:52        34296   8129800     99.58     43184   6524872    976776     75472      7.17     18340
21:50:55        52336   8111760     99.36     43188   6505392    976776     75472      7.17     18340
21:50:58        51740   8112356     99.37     43196   6505384    976776     75472      7.17     18340
Average:        46124   8117972     99.44     43189   6511883    976776     75472      7.17     18340
```

- 监控网卡流量

```
ZKY_1:~ # sar -n DEV 2 2
Linux 2.6.5-7.287.3-smp (ZKY_1)          11/24/09

21:52:05        IFACE   rxpck/s   txpck/s   rxbyt/s   txbyt/s   rxcmp/s   txcmp/s  rxmcst/s
21:52:07           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
21:52:07     priveth0      1.01      1.01     78.39     78.39      0.00      0.00      0.00
21:52:07     priveth1     20.10     22.11   4240.20   4616.08      0.00      0.00      0.00
21:52:07      pubeth0      3.52      2.51    285.43    260.30      0.00      0.00      0.00
21:52:07      pubeth1     72.86     91.46  16537.69  50510.55      0.00      0.00      0.00
21:52:07      pubeth2      1.51      3.02     96.48    192.96      0.00      0.00      0.00
21:52:07      pubeth3      1.01      3.02     64.32    192.96      0.00      0.00      0.00
21:52:07         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
21:52:07        bond0     75.38     97.49  16698.49  50896.48      0.00      0.00      0.00
```

性能调优检查项——CPU利用率

```
top - 21:58:16 up 3 days, 21:57,  1 user,  load average: 1.12, 1.16, 1.18
Tasks: 294 total,   1 running, 292 sleeping,   0 stopped,   1 zombie
Cpu(s):  0.6% us,  3.6% sy,  0.0% ni, 93.7% id,  0.8% wa,  0.2% hi,  1.1% si
Mem:   8164096k total,  7714736k used,   449360k free,    39908k buffers
Swap:  1052248k total,    75464k used,   976784k free,  6114792k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
```

```
top - 14:48:17 up 4 days, 13 min,  8 users,  load average: 18.73, 18.81, 18.62
Tasks: 433 total,   2 running, 430 sleeping,   1 stopped,   0 zombie
Cpu0  :  5.0%us,  1.3%sy,  0.0%ni, 93.7%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu1  :  1.0%us,  1.3%sy,  0.0%ni, 97.7%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu2  :  0.0%us,  0.3%sy,  0.0%ni, 99.0%id,  0.7%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu3  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu4  :  3.3%us,  2.6%sy,  0.0%ni, 94.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu5  :  2.6%us,  2.0%sy,  0.0%ni, 95.4%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu6  :  0.7%us,  0.3%sy,  0.0%ni, 99.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu7  :  0.0%us,  0.3%sy,  0.0%ni, 99.7%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:    193128M total,   19577M used,   173550M free,      497M buffers
Swap:     2596M total,       0M used,     2596M free,     1359M cached
```


HUAWEI

性能调优检查项-iostat命令

```
Time: 19:32:12
avg-cpu:  %user   %nice    %sys %iowait   %idle
           0.00    0.00    0.00    0.00  100.00

Device:    rrqm/s wrqm/s   r/s   w/s  rsec/s  wsec/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
hda          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdb          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdc          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdd          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sde          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdf          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdg          0.00   0.00  0.00  0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
```


HUAWEI

常用的性能分析命令

- Top：显示当前进程的工作状态。
- Vmstat：显示当前Linux系统的工作状态。
- netstat
  - netstat –s，显示TCP统计信息；
  - netstat –t，显示每一条TCP连接的状态；
- Iostat：显示磁盘IO当前状态，以及当前每个盘的利用率和吞吐量。
- Iperf：测试网络吞吐量。

 Page 36 HUAWEI

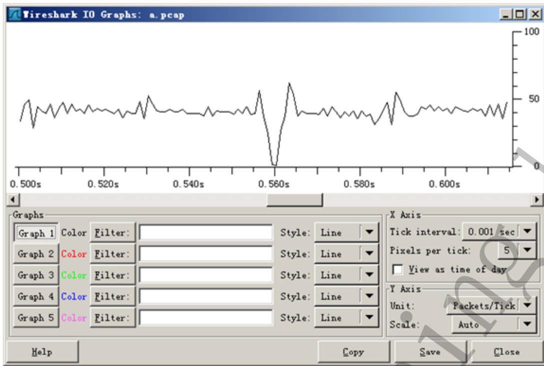
在这里列出每一个在网络性能调整过程中可能用到的工具，仅仅起到参考作用。此处并没有详细介绍每个工具的具体用法，详细用法参见每个工具的参考手册。

常用的性能分析和调整工具（一）

- ethtool：显示网卡状态，配置网卡参数
  - ethtool ifname，显示网卡的连接速率和双工状态
  - ethtool －c ifname，显示网卡的中断聚合配置
  - ethtool －C ifname，更改网卡的中断聚合配置
  - ethtool －g ifname，显示网卡接收buffer和发送buffer的数量
  - ethtool － G ifname，更改网卡接收buffer和发送buffer的数量
  - ethtool －k ifname，显示网卡的TOE配置
  - ethtool －K ifname，更改网卡的TOE配置
  - ethtool －a ifname，显示网卡的流控配置
  - ethtool －A ifname，更改网卡的流控配置

常用的性能分析和调整工具（二）

- iptables
  - 显示和配置Linux防火墙
- tcpdump，wireshark



  HUAWEI

tcpdump，wireshark是抓包工具，在Linux上可以用tcpdump eth0 –w out.pcap把报文存储下来，然后用wireshark打开这个文件来分析。

常用的性能分析和调整工具（三）

- /proc/interrupts
  - 显示当前每个CPU核处理中断的数量
- Taskset
  - 设置把进程邦定到特定CPU核上
- /proc/irq/xxx/smp_affinity
  - 把中断邦定到特定CPU核上
- sar
  - sar －n DEV
    - 显示当前网络卡的吞吐量
  - sar －n NFSD
    - 显示当前nfsd服务的请求数量



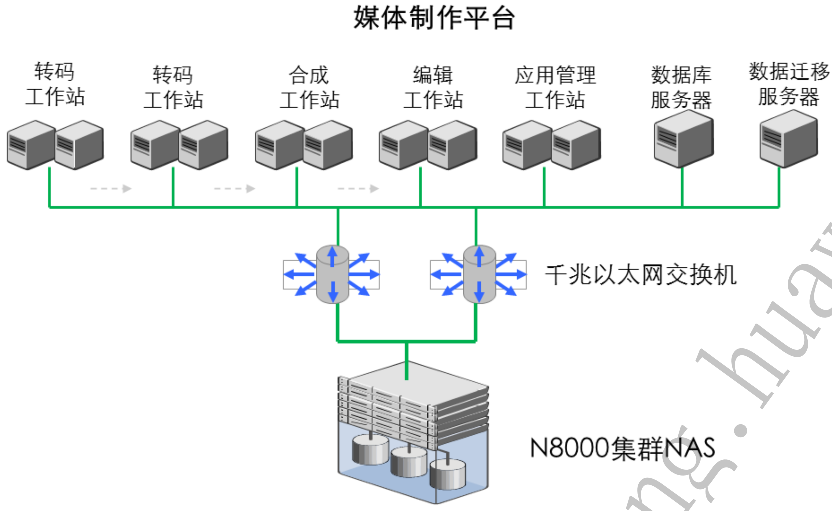
Oprofile是系统性能剖析工具，能够动态分析影响性能的各种因素，包括：

- 程序运行过程中间消耗CPU性能最多的代码片段
- CPU内部cache的命中率
- CPU内部TLB的命中率
- CPU内存接口带宽的利用率

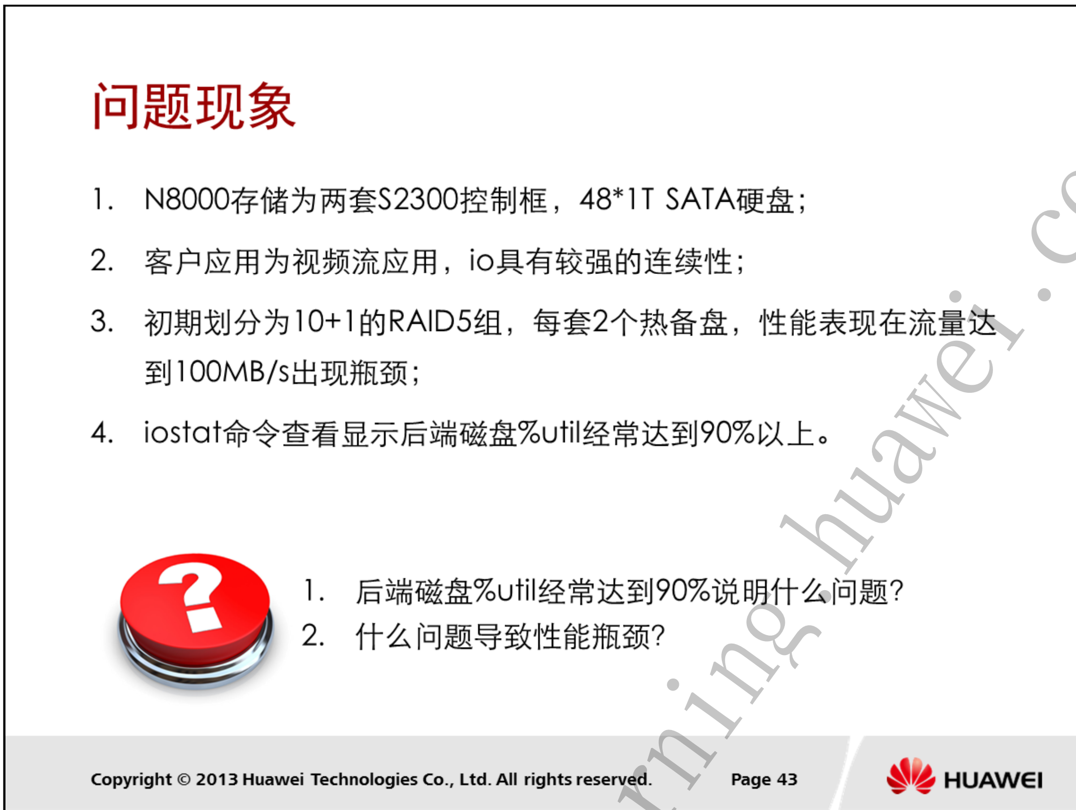
常用的性能分析和调整工具（五）

- numastat
  - 显示“非对称多路内存访问”机器的NUMA工作状态。
- mpstat
  - 显示多核CPU的利用率。
- free
  - 显示剩余内存的数量。
- ionice、nice
  - ionice调整进程的io优先级，nice调整进程的CPU调度优先级。

性能优化案例分享：某媒资应用



 Page 42 HUAWEI



- 问题分析：
  - 使用sar，iostat，top命令查看N8000节点cpu、内存利用率，空闲率较高，节点上不存在瓶颈；属于读写混合性业务，读写比率为1：2左右
  - 通过cm shilo命令观察Io大小以及连续性，读IO大小为32K，写IO最大为256K
  - iostat查看存储单元读写，在纯写业务时磁盘上同时有大量的读操作
  - 创建两个文件系统，每个文件系统的columns数目设置为4，对应两个raid组，不利于IO性能提升
  - LUN条带单元大小设置为64K，整个条带大小为640K，存在大量的写惩罚

配置优化调整

- 配置调整措施:
  1. 重新划分RAID组，创建5+1的raid 5组，每套设置1个热备盘
  2. 重新创建文件系统，创建文件系统列数为4，每个lun取自一个RAID组
  3. 重新设置LUN条带单元大小为32K
  4. cache水位设置为85/65
  5. 预取策略设置为异步预取
- 配置优化结果:
  - 新配置完成后，继续观察业务运行情况，写惩罚大大降低，iowait延迟降低，磁盘利用率降低，性能达到客户要求，问题解决

总结

- N8500整体规划思路
- 网络规划
- 容量与文件系统规划
- 文件共享业务规划
- 性能问题排查方法和优化建议

思考题

- N8500 物理IP、虚拟IP、管理IP规划时需要注意哪些?
- RAID组、LUN和文件系统等空间规划，有哪些最佳建议?

习题

1. （单选题）SATA、NL-SAS盘推荐配置为（　）

   A. 11盘RAID 6； B　12 盘 RAID5； C　2盘RAID10； D　4盘RAID5

2. （判断题）N8500所有IP地址（Console IP地址、物理IP地址、业务网口虚拟IP地址）在同一网段内。（ ）

3. （多选题）N8500系统有关LUN规划，关键参数最佳建议描述正确的有（　）

   A. 单个LUN推荐最大不超过16TB

   B. 同一RAID组的LUN归属于同一控制器，不同的RAID组的LUN交替归属于A控和B控。

   C. 写cache：采用回写不镜像

   D. LUN映射给集群主机时，主机LUN ID不能为0

Page 47　HUAWEI

- 1、A；2、T；3 ABD

Thank you

www.huawei.com

HC1209208 大数据存储系统规划与设计

HC1209208
第八章 大数据存储规划与设计

www.huawei.com

HUAWEI

目标

- 学完本课程后，您将能够:
  - 掌握OceanStor 9000 前端GE后端10GE组网规划
  - 掌握OceanStor 9000 IB组网规划
  - 掌握OceanStor 9000 GE组网规划

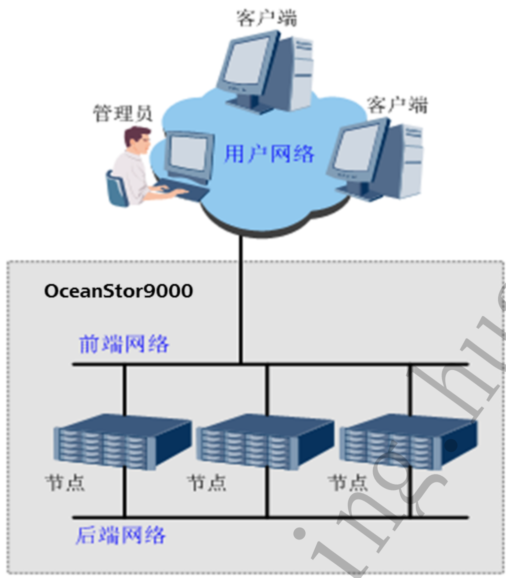
目录



- OceanStor 9000分为以下三类节点：
  - Mini-Capacity node:2U 12盘位存储节点，应用于小容量应用场景（Mini-Capacity appplications）
  - Performance node: 2U 25盘位存储节点，应用于OPS密集型应用场景（OPS-intensive applications）
  - Capacity node: 4U 36盘位存储节点，应用于大容量、高带宽应用场景（Capacity applications）
- OceanStor 9000组网使用的交换机
  - GE交换机：S5700-52C-EI，简称S5752,也可使用客户现网GE交换机
  - 10GE交换机：CE6850-48S4Q-EI，简称CE6850
  - IB交换机：IS5023，也可使用客户现网前端业务交换机

9000组网构成

- 前端业务网络
- 后端存储网络
- IPMI管理网络



Page 4
HUAWEI

前端业务网络是OceanStor 9000与用户网络对接的组网，用于客户端接入业务和登录OceanStor 9000管理控制台，简称前端网络。后端存储网络是OceanStor 9000内部节点间的组网，用于集群内部通信，简称后端网络。IPMI管理网络是连接节点IPMI接口，用于登录节点的IPMI管理平台，简称IPMI网络。

9000部署方案

- 10GE组网部署方案
- IB组网部署方案
- GE组网部署方案

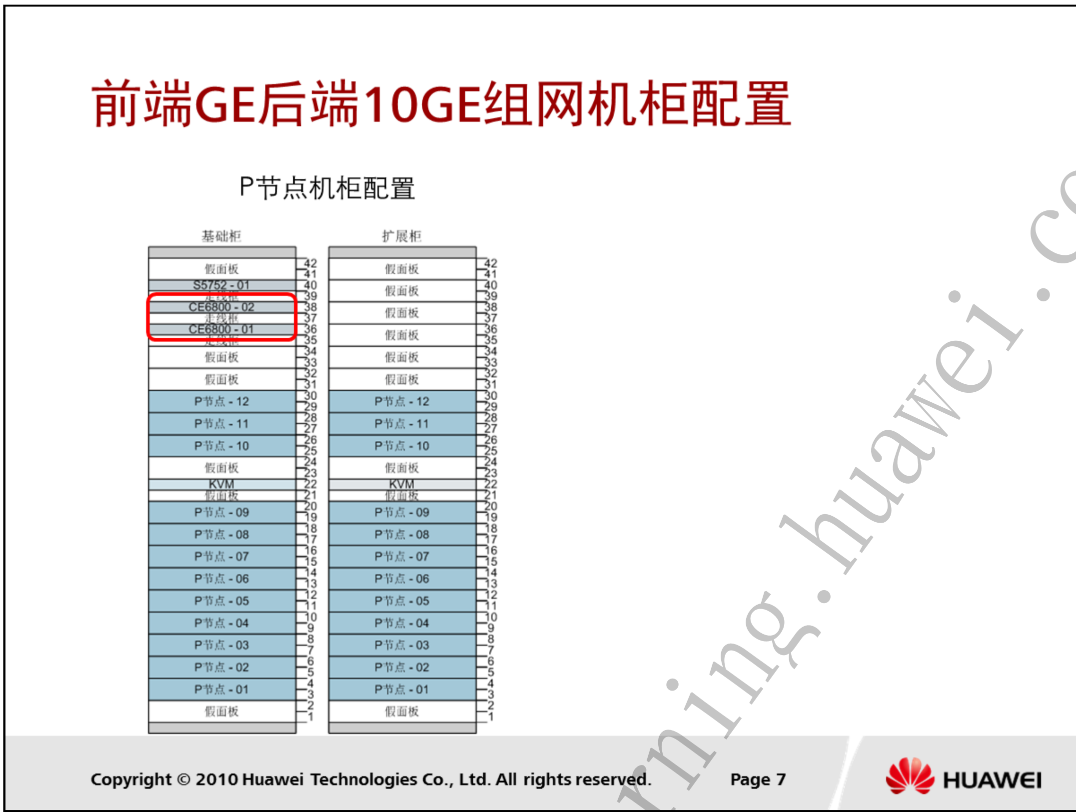
方案	前端业务网络接入方式	后端存储网络接入方式
10GE组网部署方案	10GE/GE	10GE
IB组网部署方案	10GE或IB	IB
GE组网部署方案	GE	GE

 Page 5 HUAWEI

- 9000节点根据不同的组网方案分为以下部署方案：
  - 前端GE后端10GE组网部署方案
  - IB组网部署方案
  - GE组网部署方案

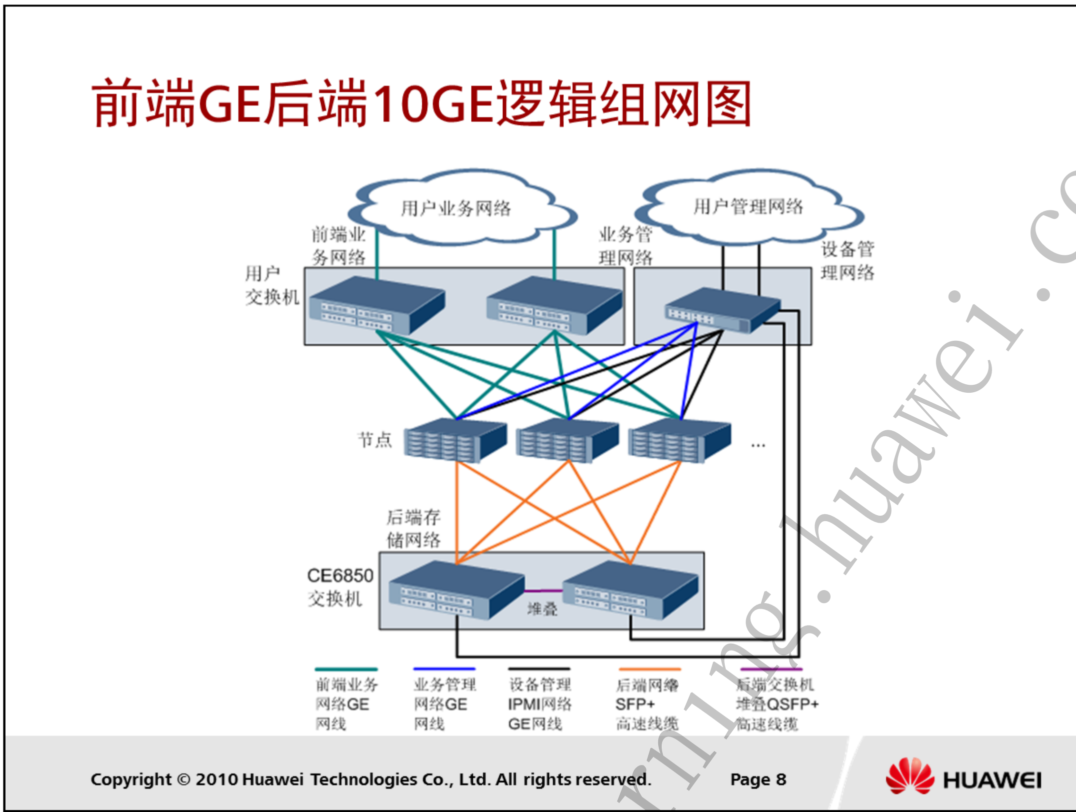
目录

HUAWEI

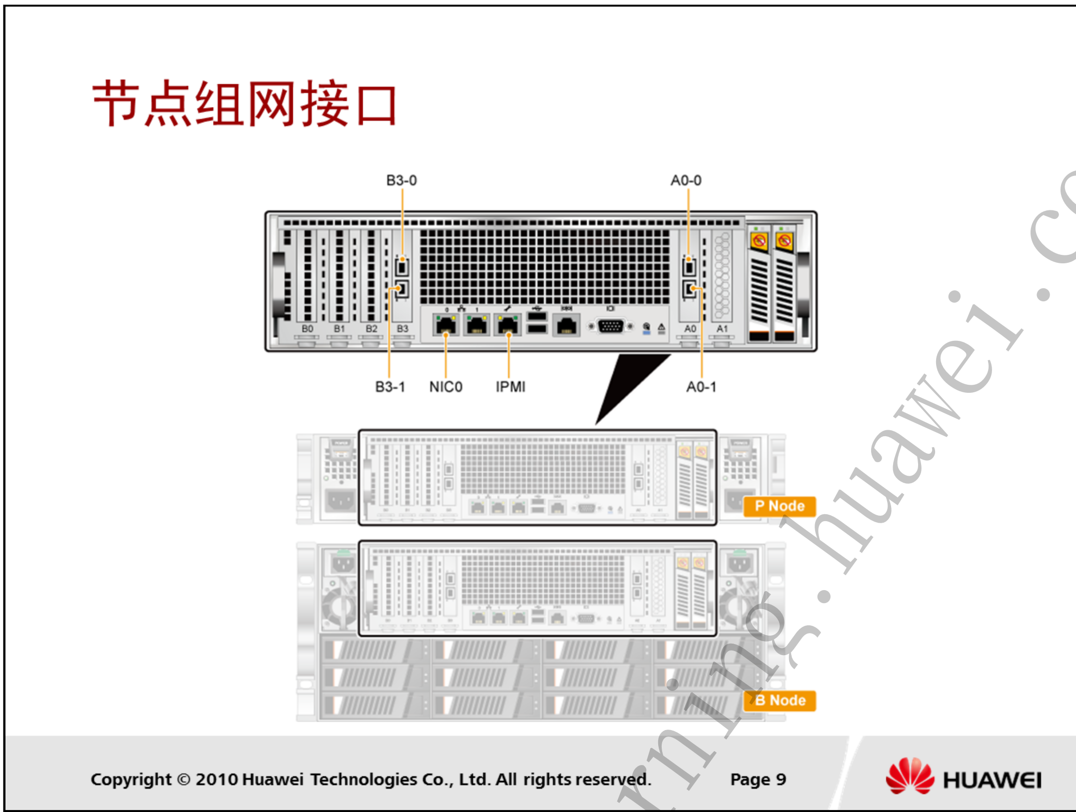


P节点（2U）3节点、5节点和C节点（4U）4节点部署仅需要配置基础柜，不需要配置扩展柜。 C节点9节点部署需要配置基础柜和扩展柜，扩展柜内部的节点通过网线和SFP+高速线缆（或光纤），连接到基础柜中的交换机。

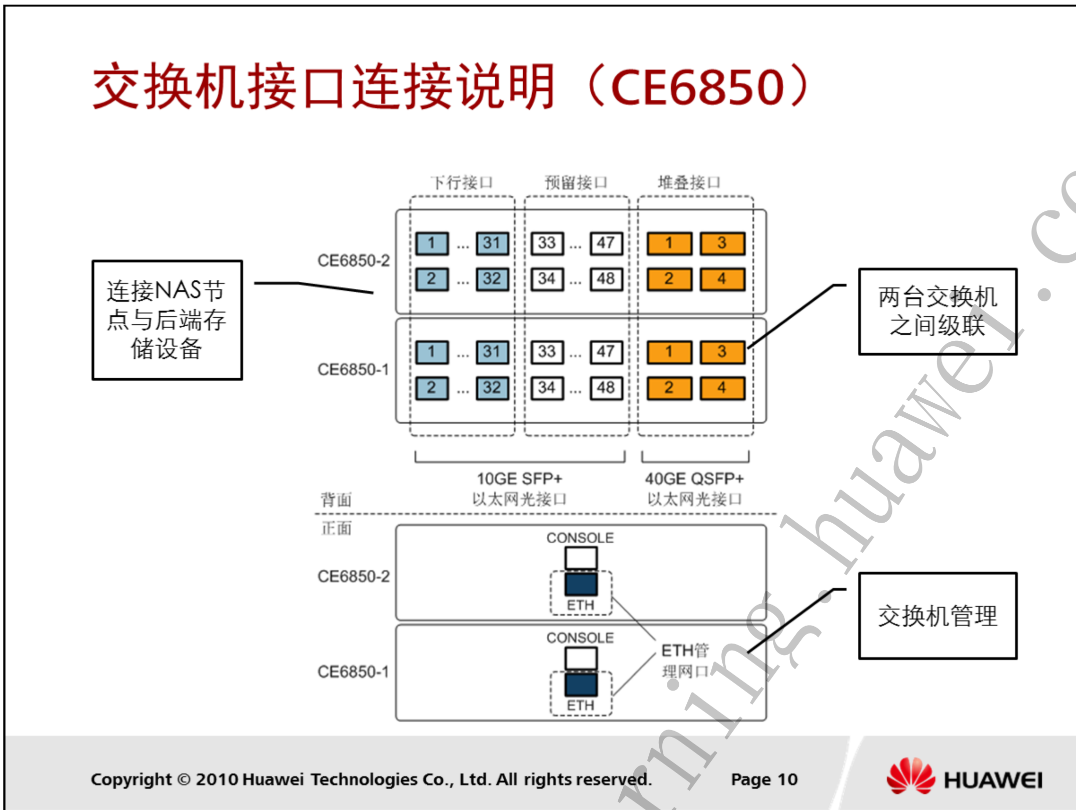
初始部署时如果节点数超过24时，为了方便后续扩容，建议使用CE12800系列交换机替换CE6850交换机，CE12800交换机安装在独立的交换机机柜中，OceanStor 9000机柜中CE6850交换机位置替换为假面板。



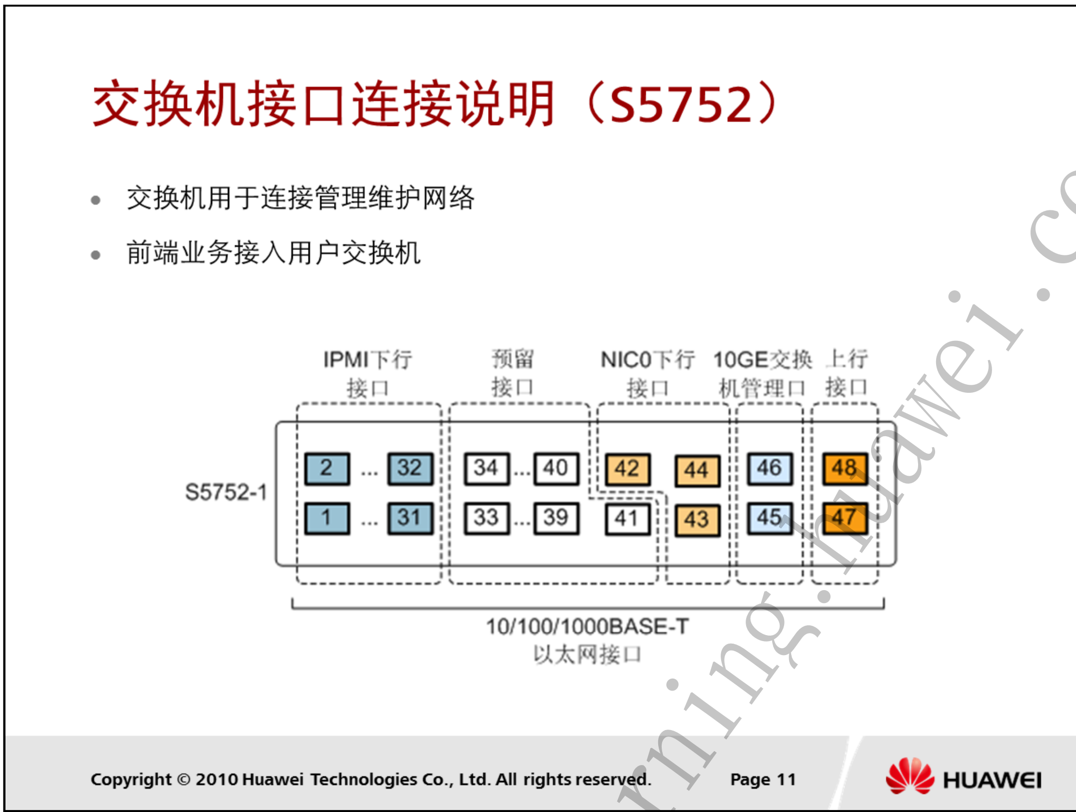
该部署方案适用于前端业务网络接入用户交换机，当前端业务网络和后端存储网络都接入CE6850交换机，此时CE6850交换机需要划分VLAN区分前端业务网络和后端存储网络。



- A0-0和A0-1（10GE光接口）
  - 连接到CE6850交换机。需要为两个接口分别规划1个后端网络IP地址。
- B3-0和B3-1（GE电接口）
  - 连接到用户前端业务交换机。需要为两个接口分别规划1个前端网络IP地址，并且为每个节点同时规划2个动态前端IP地址。
- NIC0（GE电接口）
  - OceanStor 9000集群中前3个节点提供管理维护接口，这些接口连接到客户管理维护交换机。无需为这些接口单独分配IP地址，部署时为集群配置的浮动管理IP地址将自动在这些接口上浮动。
- IPMI（FE电接口）
  - 连接到客户管理维护交换机。需要为该接口规划1个IP地址，连接到用户管理维护网络。



- 下行接口（连接NAS节点）
  - CE6850-1连接各节点A0-0接口。
  - CE6850-2连接各节点A0-1接口。
- 堆叠接口（交换机间互联）①
  - 节点数量≤16，使用2个接口。
  - 16＜节点数量≤24，使用3个接口。
  - 24＜节点数量≤32，使用4个接口。
- 预留接口
  - 空闲接口，建议执行shutdown命令，关闭这些接口。
- ETH管理网口
  - 用于交换机的管理。这两个接口连接到管理维护交换机。



- IPMI下行接口：顺序连接到各节点的IPMI接口。
- NIC0下行接口：3个接口顺序连接到集群最前面的3个节点（基础柜中01～03号节点）。
- 10GE交换机管理口：连接到两台CE6850交换机正面的ETH管理网口。
- 上行接口：连接到客户的管理维护网络。
- 预留接口：空闲接口，建议执行shutdown命令，关闭这些接口。

IP地址规划

网段类型	IP地址数量
前端业务网络	• 节点数量×6+1
后端存储网络	• 实际IP地址数量：节点数量×2 • 软件安装临时IP地址：节点数量
管理网络	• 节点数量+管理网络交换机数量+2 • 软件安装临时IP地址：1

 Page 12 HUAWEI

- 前端业务网络规划说明
  - 为每个节点分配2个IP地址，该地址为节点的静态前端IP地址。
  - 为WiseLink功能的IP地址池分配n（n=节点数量×4）个动态前端IP地址。
  - 1个WiseLink DNS IP地址。
- 后端存储网络
  - 每个存储节点分配2个后端存储网络IP地址。
  - OceanStor 9000在软件安装过程中还需要临时申请以下IP地址。
  - n（n=节点数量）个后端存储网络IP地址，软件安装时通过DHCP分配临时IP地址时，每个存储节点再分配1个。这些IP地址用于软件安装时临时分配给各节点。完成软件安装后，将自动删除这些地址。这些临时规划的后端存储网络IP地址不能与节点实际规划的后端存储网络IP地址重叠。
- 管理网络
  - 每个存储节点IPMI接口分配1个，每台管理网络交换机分配1个，OceanStor 9000管理IP地址1个，后端存储网络交换机1个。

    说明： 两台交换机配置为堆叠时，对外体现为1台交换机，此时仅需要分配1个IP地址。
  - OceanStor 9000在软件安装过程中还需要临时申请以下IP地址：
    - 1个管理网络IP地址，用于安装软件时登录部署工具并上传文件包。

IP地址规划样例（节点）

接口名称	规划示例	所属网络
A0-0	172.18.0.10	后端网络
A0-1	172.18.0.20	
Bond接口0（B3-0和B3-1）	192.168.100.10	前端网络
IPMI	10.110.100.10	管理维护网络



规划项目	规划示例	说明	所属网络
后端网络掩码	掩码：255.255.252.0	后端网络用于OceanStor 9000内部通信，无需规划网关地址，部署时需要按实际需求规划掩码。	后端网络
前端网络网关地址和掩码	IP地址：192.168.100.1 掩码：255.255.255.0	前端网络用于OceanStor 9000与客户端通信。	前端网络
WiseLink DNS IP地址	192.168.100.100	当部署WiseLink特性时，需要为OceanStor 9000规划动态前端IP地址和WiseLink DNS IP地址。	
动态前端IP地址	192.168.100.110～192.168.100.150		
OM管理IP	192.168.100.200	管理员可以使用该IP地址登录OceanStor 9000管理控制台。该IP地址为浮动IP地址，可以浮动到集群内所有节点的前端网络接口。	
管理网络网关地址和掩码	IP地址：10.110.100.1 掩码：255.255.255.0	用于OceanStor 9000节点与客户维护管理系统通信。	管理维护网络

当部署WiseLink特性时，需要为OceanStor 9000规划动态前端IP地址和WiseLink DNS IP地址。其中动态前端IP地址需要在完成OceanStor 9000软件安装后，在OceanStor 9000管理控制台上配置。

IP地址规划样例（配置文件样例）

- 前3个节点使用NIC0接口连接管理网络，需要独立出管理网口

部署节点				
节点SN	210235G78MZ0CC000031	节点名称	Node1	节点池
IPMI IP地址	10.110.100.10	机架号	1	槽位号
网口是否bond			No	
后端IP地址1	172.18.0.10	后端IP地址1的网口	A0-0	
后端IP地址2	172.18.0.20	后端IP地址2的网口	A0-1	
前端IP地址1	192.168.100.10	前端IP地址1的网口	B3-0	
前端IP地址2	192.168.100.20	前端IP地址2的网口	B3-1	
是否独立出管理网口	Yes	OM管理网口	NIC0	

- 从第4个节点开始将不连接NIC0接口到管理网络。

部署节点				
节点SN	210235G78MZ0CC000032	节点名称	Node4	节点池
IPMI IP地址		机架号	1	槽位号
网口是否bond			No	
后端IP地址1	172.18.0.13	后端IP地址1的网口	A0-0	
后端IP地址2	172.18.0.23	后端IP地址2的网口	A0-1	
前端IP地址1	192.168.100.13	前端IP地址1的网口	B3-0	
前端IP地址2	192.168.100.23	前端IP地址2的网口	B3-1	
是否独立出管理网口	No	OM管理网口		

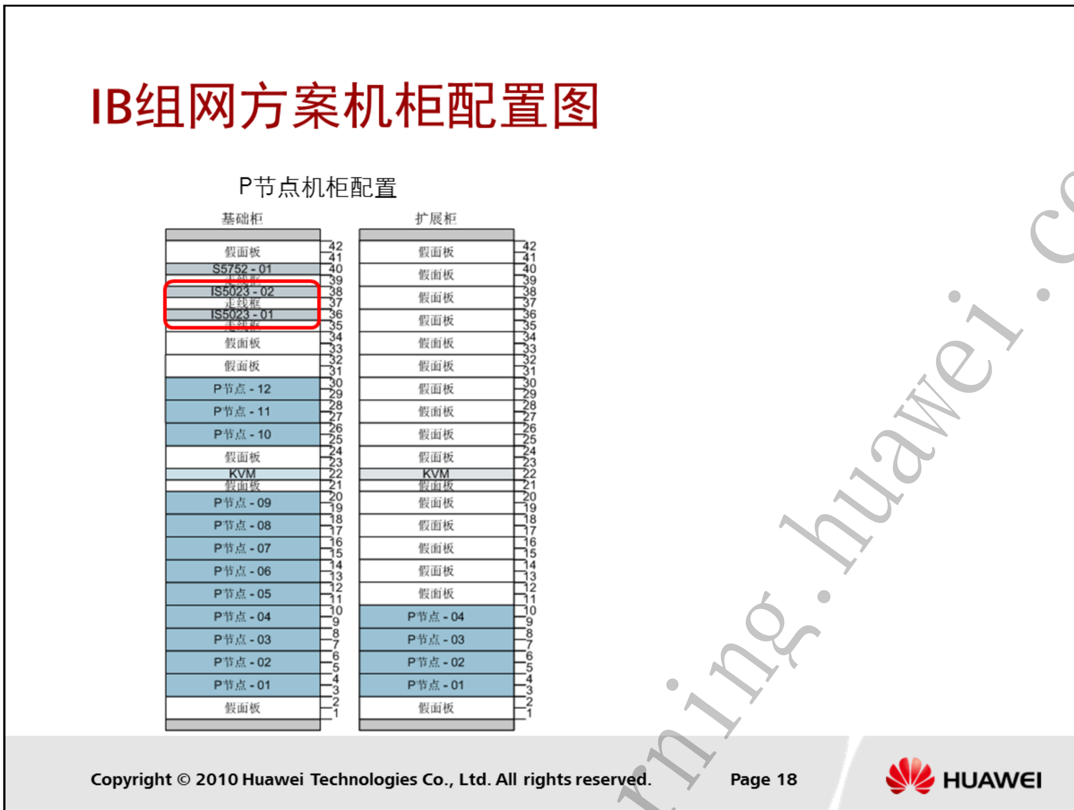
交换机配置说明

- CE6850交换机配置说明
  - 配置交换机管理IP
  - 配置两台交换机堆叠
  - 配置SNMP
  - 配置Telnet登录和Console登录
- S5752交换机配置说明
  - 配置交换机管理IP
  - 配置SNMP
  - 配置Telnet登录和Console登录

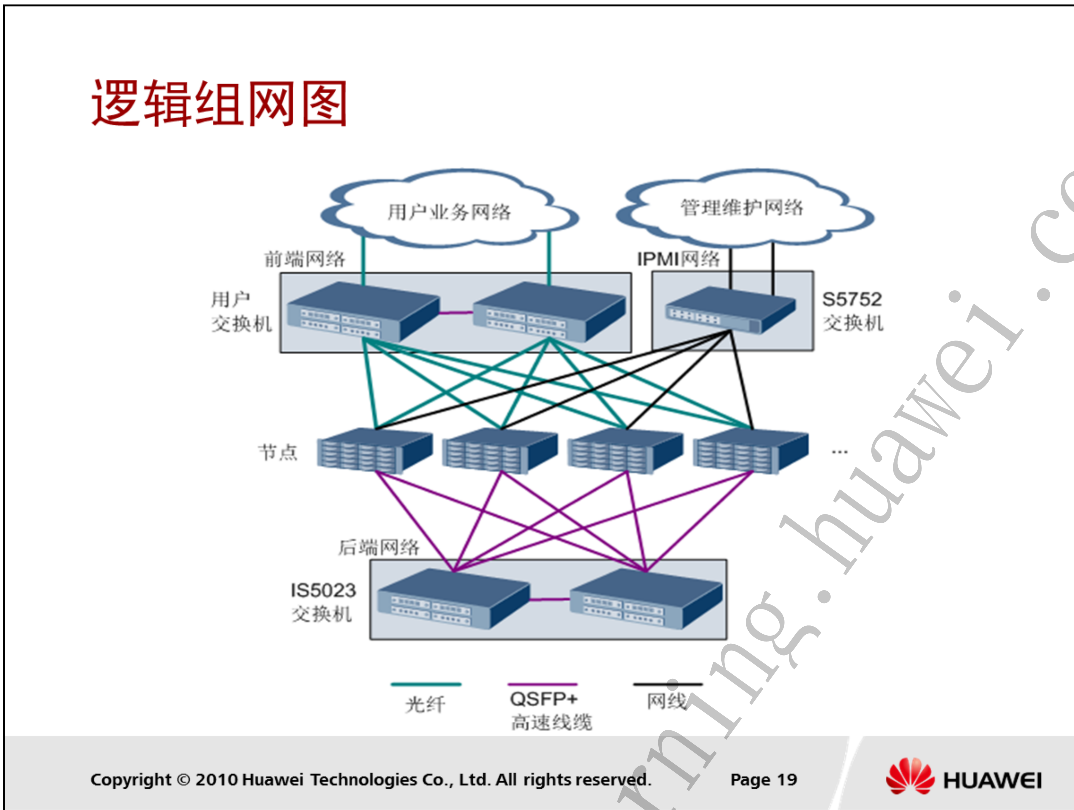
HUAWEI

目录

HUAWEI

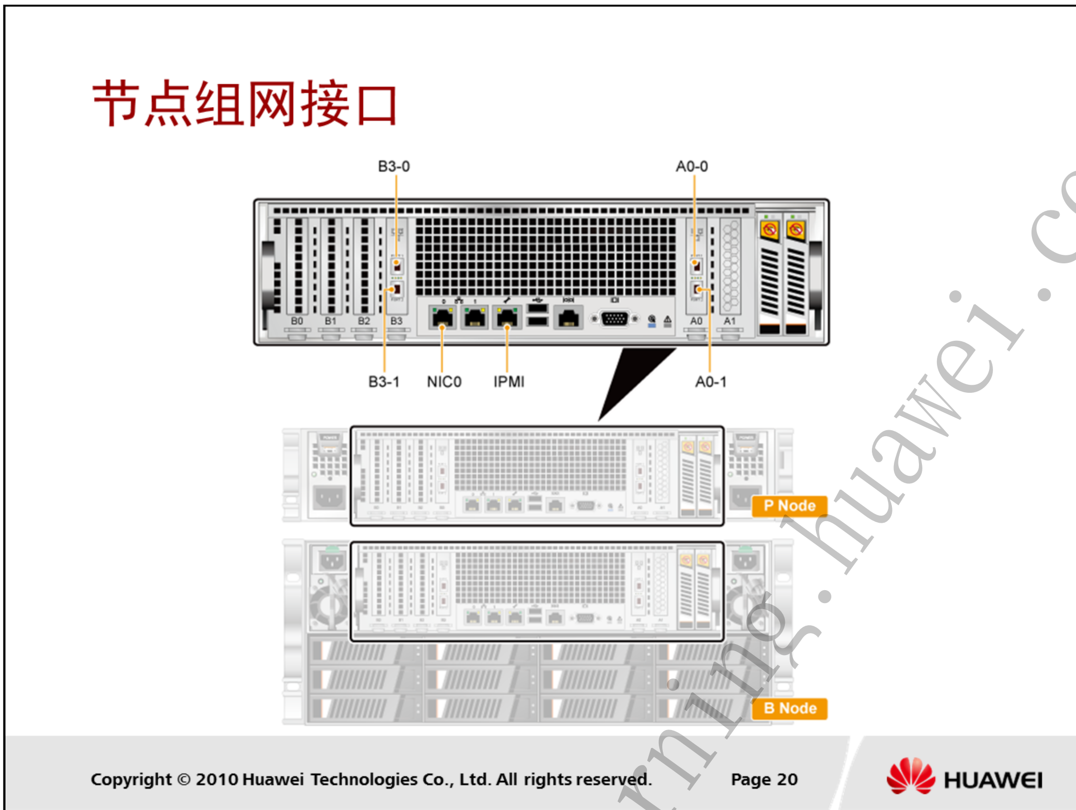


扩展柜内部的节点需要通过网线和QSFP+高速线缆（或光纤），连接到基础柜中的交换机。初始部署时如果节点数超过16时，则需要根据实际组网规模，确定IB交换机的型号。



- 节点数量不超过16个时，采用上图典型组网方案，该方案配置如下：
  - P节点机柜配置为1个基础柜加上1个扩展柜，其中扩展柜最多配置4个节点。
  - C节点机柜配置为1个基础柜加上1个扩展柜。

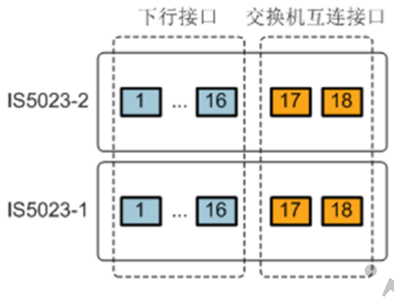
本组网方案中，后端网络采用IB接口，前端业务网络接入用户交换机。用户接入交换机可采用10GB接口和IB接口。



- A0-0和A0-1（IB接口）
  - 分别连接到两台IS5023交换机。需要为两个接口分别规划1个后端网络IP地址。
- B3-0和B3-1（IB接口或10GE光接口）
  - 连接到用户的交换机。
  - 采用IB接口时，需要为两个接口分别规划1个前端网络IP地址。
  - 采用10GE光接口时，如果用户交换机配置为堆叠，建议两个接口配置为Bond接口，接口模式为Bond6，需要为该Bond接口规划1个前端网络IP地址。当用户交换机未配置堆叠时，需要为两个接口分别规划1个前端网络IP地址。
- NIC0（GE电接口）
  - OceanStor 9000集群中前3个节点提供管理维护接口，这些接口连接到S5752交换机。
  - 无需为这些接口单独分配IP地址，部署时为集群配置的管理IP地址将自动在这些接口上浮动。
- IPMI（FE电接口）
  - 连接到S5752交换机。需要为该接口规划1个IP地址，连接到用户管理维护网络。

交换机接口连接说明（IS5023）

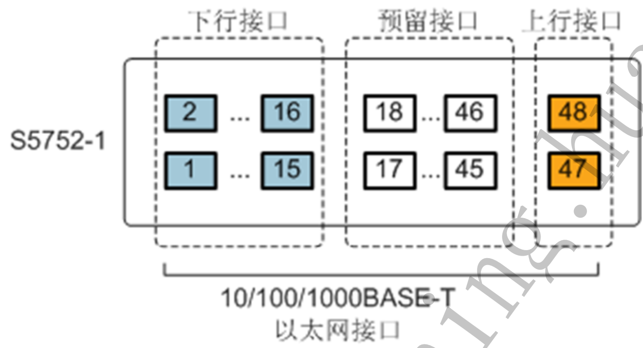
- 下行接口
  - IS5023-1交换机连接各节点A0-0接口，IS5023-2交换机连接各节点A0-1接口。
- 交换机互连接口
  - 用于两台IS5023交换机互连。



HUAWEI

交换机接口连接说明（S5752）

- 下行接口: 顺序连接到各节点的IPMI接口。
- 上行接口:连接到客户的管理维护网络。
- 预留接口:空闲接口，建议执行shutdown命令，关闭这些接口。



HUAWEI

IP地址规划

网段类型	IP地址数量
前端业务网络	• 节点数量×6+1
后端存储网络	• 实际IP地址数量：节点数量×2 • 软件安装临时IP地址：节点数量
管理网络	• 节点数量+管理网络交换机数量+1 • 软件安装临时IP地址：1

 Page 23 HUAWEI

- 前端业务网络规划说明
  - 为每个节点分配2个IP地址，该地址为节点的静态前端IP地址。
  - 为WiseLink功能的IP地址池分配n（n=节点数量×4）个动态前端IP地址。
  - 1个WiseLink DNS IP地址。
- 后端存储网络
  - 每个存储节点分配2个后端存储网络IP地址。
  - OceanStor 9000在软件安装过程中还需要临时申请以下IP地址。
    - n（n=节点数量）个后端存储网络IP地址，软件安装时通过DHCP分配临时IP地址时，每个存储节点再分配1个。这些IP地址用于软件安装时临时分配给各节点。完成软件安装后，将自动删除这些地址。这些临时规划的后端存储网络IP地址不能与节点实际规划的后端存储网络IP地址重叠。
- 管理网络
  - 每个存储节点IPMI接口分配1个，每台管理网络交换机分配1个，OceanStor 9000管理IP地址1个。
  - OceanStor 9000在软件安装过程中还需要临时申请以下IP地址：
  - 1个管理网络IP地址，用于安装软件时登录部署工具并上传文件包。

IP地址规划样例（节点）

接口名称	规划示例	所属网络
A0-0	172.18.0.10	后端网络
A0-1	172.18.0.20	
B3-0	192.168.100.10	前端网络
B3-1	192.168.100.20	
IPMI	10.110.100.10	管理维护网络

 Page 24 HUAWEI

本实例以用户前端业务网络交换机使用10GE交换机为例说明。当用户的前端业务网络交换机使用IB交换机时，节点使用接口变化如下：

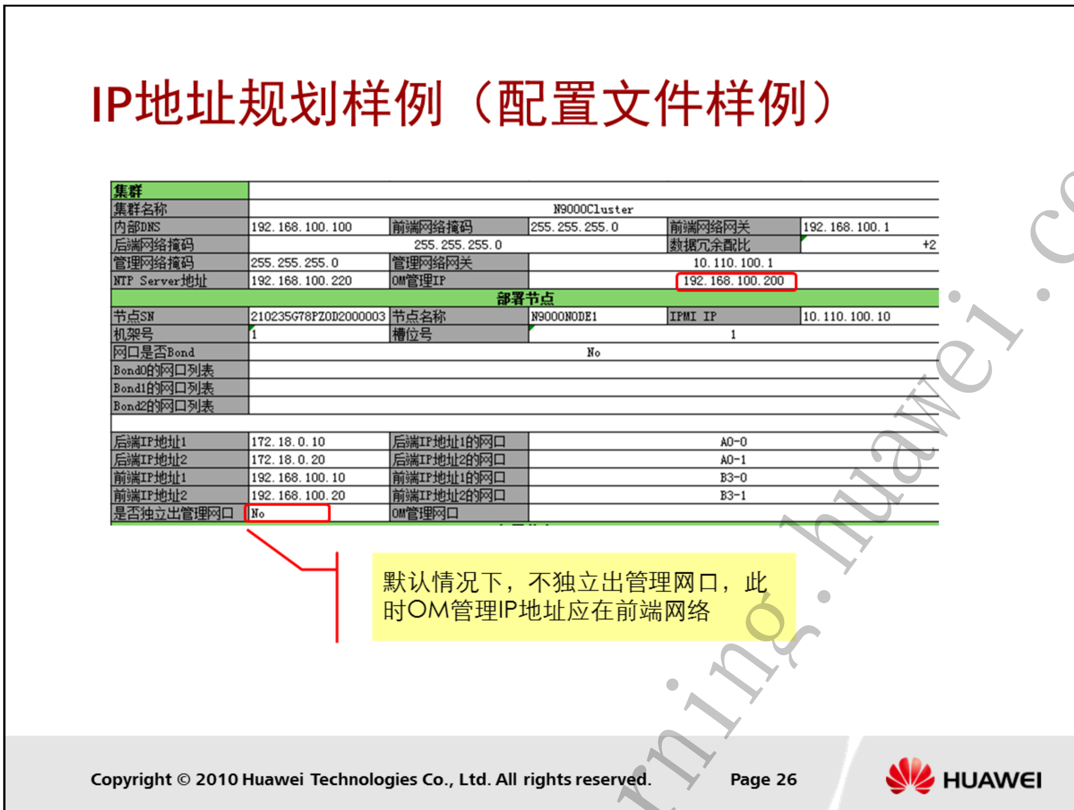
- A0-0和B3-0连接后端存储网络交换机。
- A0-1和B3-1连接前端业务网络交换机。

IP地址规划样例（集群）

规划项目	规划示例	说明	所属网络
后端网络	掩码：255.255.252.0	后端网络用于OceanStor 9000内部通信，无需规划网关地址，部署时需要按实际需求规划掩码。	后端网络
前端网络网关地址	IP地址：192.168.100.1 掩码：255.255.255.0	前端网络用于OceanStor 9000与客户端通信。	前端网络
WiseLink DNS IP	192.168.100.100	当部署WiseLink特性时，需要为OceanStor 9000规划动态前端IP地址和WiseLink DNS IP地址。	
动态前端IP	192.168.100.110～192.168.100.150		
OM管理IP	192.168.100.200	管理员可以使用该IP地址登录OceanStor 9000管理控制台，该IP地址为浮动IP地址。该地址可以浮动到集群内所有节点的前端网络接口。	
管理网络网关地址	IP地址：10.110.100.1 掩码：255.255.255.0	用于OceanStor 9000节点与客户维护管理系统通信。	管理维护网络

 Page 25 HUAWEI

当部署WiseLink特性时，需要为OceanStor 9000规划动态前端IP地址和WiseLink DNS IP地址。其中动态前端IP地址需要在完成OceanStor 9000软件安装后，在OceanStor 9000管理控制台上配置。



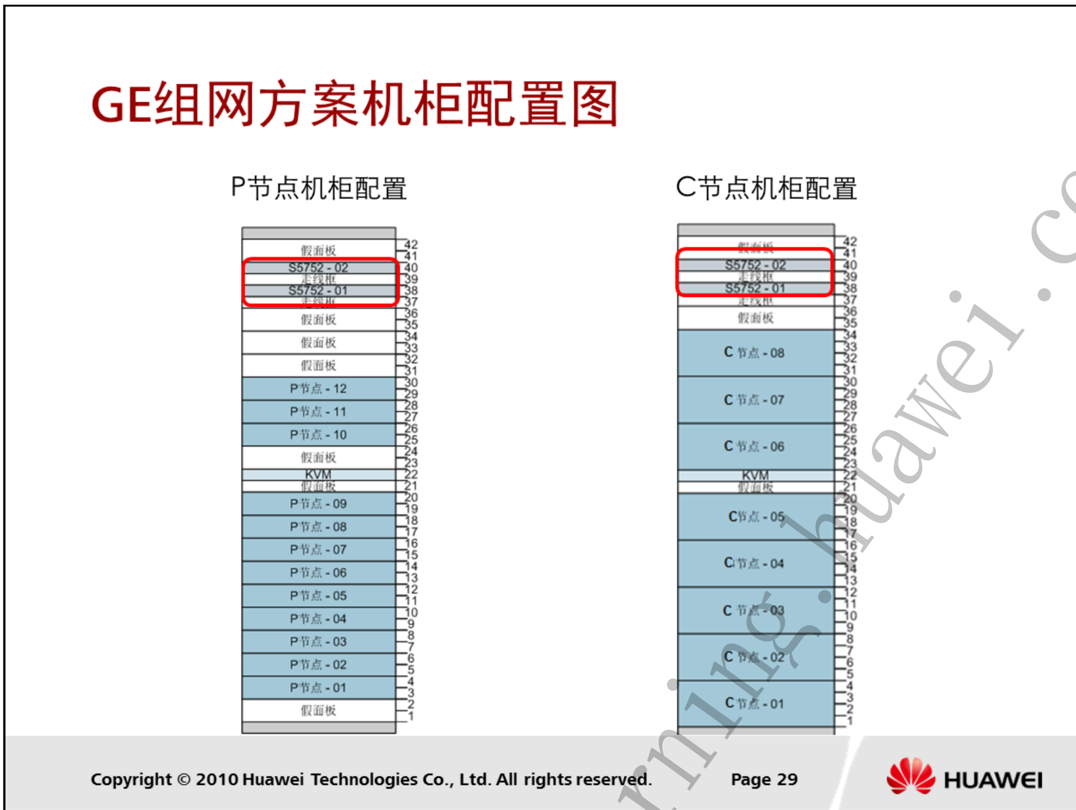
集群					
集群名称	N9000Cluster				
内部DNS	192.168.100.100	前端网络掩码	255.255.255.0	前端网络网关	192.168.100.1
后端网络掩码	255.255.255.0			数据冗余配比	+2
管理网络掩码	255.255.255.0	管理网络网关	10.110.100.1		
NTP Server地址	192.168.100.220	OM管理IP	192.168.100.200		
部署节点					
节点SN	210235G78PZ0D2000003	节点名称	N9000NODE1	IPMI IP	10.110.100.10
机架号	1	槽位号	1		
网口是否Bond	No				
Bond0的网口列表					
Bond1的网口列表					
Bond2的网口列表					
后端IP地址1	172.18.0.10	后端IP地址1的网口	A0-0		
后端IP地址2	172.18.0.20	后端IP地址2的网口	A0-1		
前端IP地址1	192.168.100.10	前端IP地址1的网口	B3-0		
前端IP地址2	192.168.100.20	前端IP地址2的网口	B3-1		
是否独立出管理网口	No	OM管理网口			

OceanStor 9000前3个节点使用NIC0接口连接管理网络，需要独立出管理网口，从第4个节点开始不连接NIC0接口到管理网络。集群配置时，不独立出管理网口，此时OM管理IP地址在前端网络。

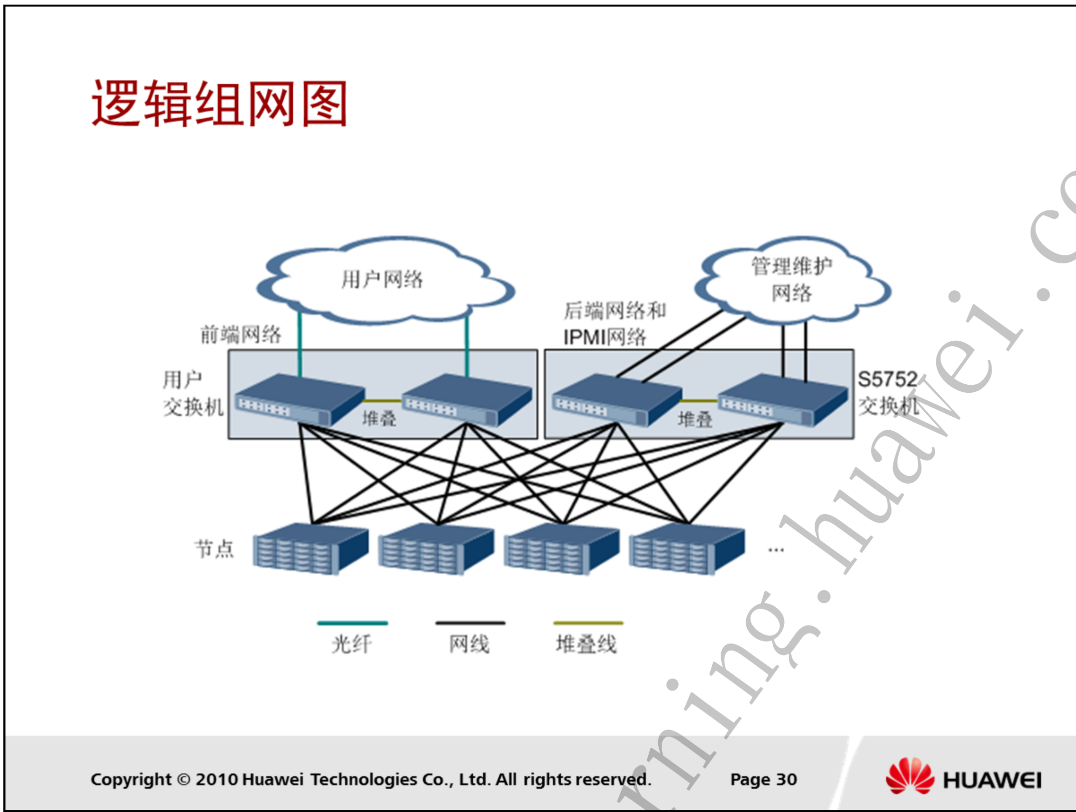
交换机配置说明

- IS5023交换机配置说明
  - IB交换机无需任何配置
- S5752交换机配置说明
  - 配置交换机管理IP
  - 配置SNMP
  - 配置Telnet登录和Console登录

目录

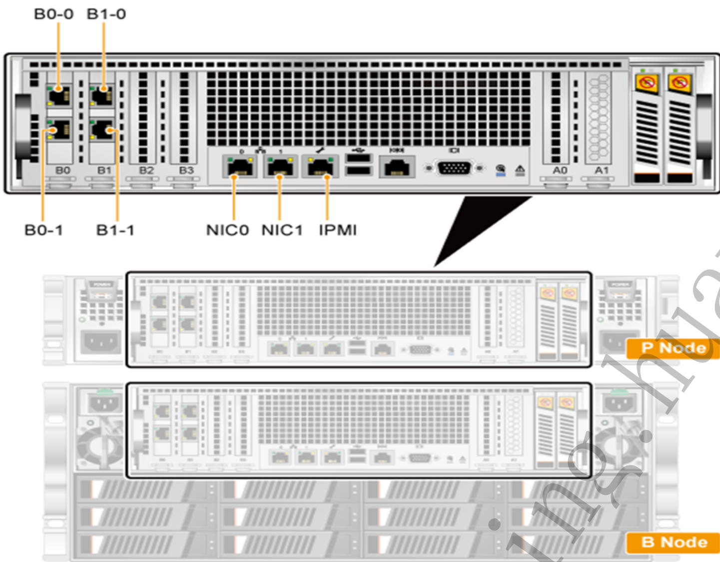


如果是多个机柜组网，需要配置两台CE6850交换机用于汇聚S5752交换机的上行连接。机柜内部两台S5752的多个上行接口需要配置链路聚合，与之对接的CE6850交换机接口也需要配置链路聚合。



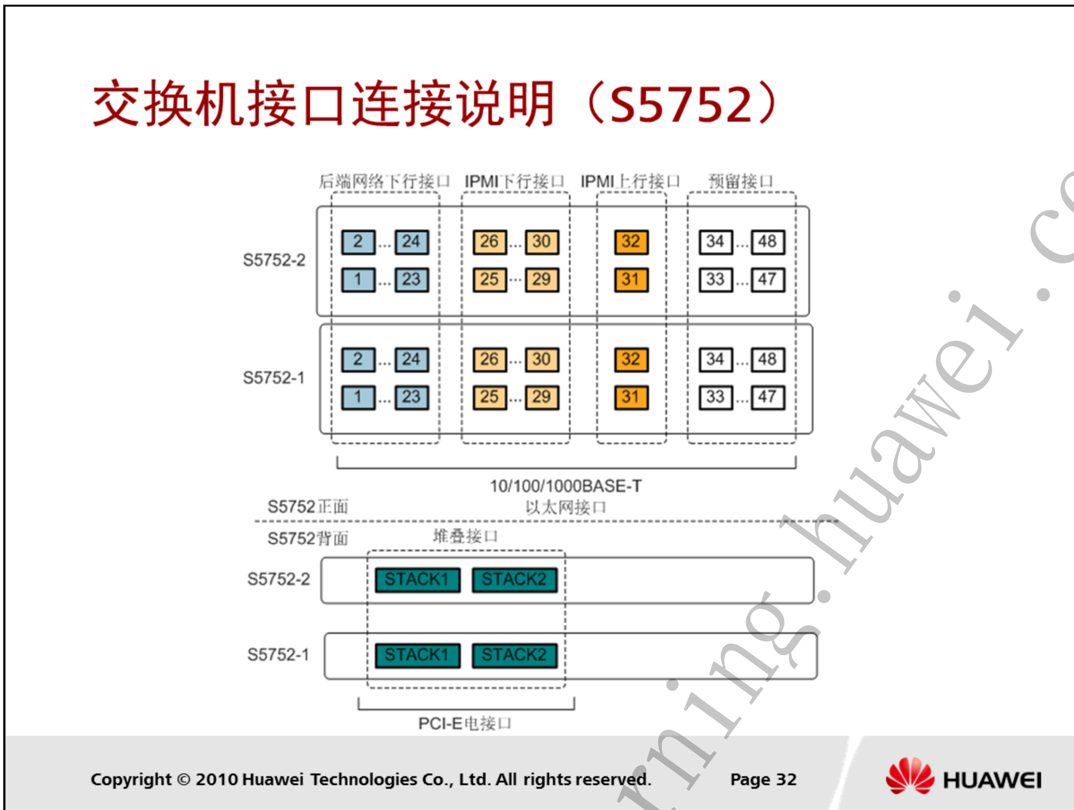
单机柜部署时采用典型组网方案，后端网络采用GE接口，前端业务网络接入用户交换机。用户接入交换机可采用GB接口。

节点组网接口



 Page 31 HUAWEI

- B0-0、B0-1、B1-0、B1-1（GE电接口）
  - B0-0和B1-0连接到S5752-01，B0-1和B1-1连接到S5752-02。4个物理接口设置为1个Bond接口，接口模式为Bond6，需要为该Bond接口规划1个后端网络IP地址。
- NIC0、NIC1（GE电接口）
  - 连接到客户的交换机。建议2个物理接口设置为1个Bond接口，接口模式为Bond6，需要为该Bond接口规划1个前端网络IP地址。
- IPMI（FE电接口）
  - 机柜内部编号为奇数的节点顺序连接到S5752-01，编号为偶数的节点顺序连接到S5752-02。需要为该接口规划1个IP地址，连接到用户管理维护网络。



- 后端网络下行接口
  - S5752-1顺序连接各个节点B0-0和B1-0。
  - S5752-2顺序连接各个节点B0-1和B1-1。
- IPMI下行接口
  - S5752-1顺序连接编号为奇数节点的IPMI接口。
  - S5752-2顺序连接编号为偶数节点的IPMI接口。
- IPMI上行接口
  - 连接到客户管理维护网络。
- 预留接口
  - 空闲接口，建议执行shutdown命令，关闭这些接口。
- 堆叠接口
  - S5752-1与S5752-2互连。STACK1连接到STACK1， STACK2连接到STACK2。

IP地址规划

网段类型	IP地址数量
前端业务网络	• 节点数量×6+2
后端存储网络	• 实际IP地址数量：节点数量×1 • 软件安装临时IP地址：节点数量
管理网络	• 节点数量+后端存储网络与管理网络交换机总数量/2

 Page 33 HUAWEI

- 前端业务网络规划说明
  - 为每个节点分配2个IP地址，该地址为节点的静态前端IP地址。
  - 为WiseLink功能的IP地址池分配n（n=节点数量×4）个动态前端IP地址。
  - 1个WiseLink DNS IP地址。
  - 1个登录OceanStor DeviceManager的管理IP。
- 后端存储网络
  - 每个存储节点分配1个后端存储网络IP地址。
  - OceanStor 9000在软件安装过程中还需要临时申请以下IP地址。
  - n（n=节点数量）个后端存储网络IP地址，软件安装时通过DHCP分配临时IP地址时，每个存储节点再分配1个。这些IP地址用于软件安装时临时分配给各节点。完成软件安装后，将自动删除这些地址。这些临时规划的后端存储网络IP地址不能与节点实际规划的后端存储网络IP地址重叠。
- 管理网络
  - 每个存储节点IPMI接口分配1个，每2台S5752交换机数量交换机分配1个。
  - 说明： 两台交换机配置为堆叠时，对外体现为1台交换机，此时仅需要分配1个IP地址。

IP地址规划样例（节点）

接口名称	规划示例	所属网络
Bond接口0 （B0-0、B0-1、B1-0、B1-1）	172.18.0.10	后端网络
Bond接口1（NIC0、NIC1）	192.168.100.10	前端网络
IPMI	10.110.100.10	管理维护网络

IP地址规划样例（集群）

规划项目	规划示例	说明	所属网络
后端网络掩码	掩码：255.255.252.0	后端网络用于OceanStor 9000内部通信，无需规划网关地址，部署时需要按实际需求规划掩码。	后端网络
前端网络网关地址	IP地址：192.168.100.1 掩码：255.255.255.0	前端网络用于OceanStor 9000与客户端通信。	前端网络
WiseLink DNS IP	192.168.100.100	当部署WiseLink特性时，需要规划动态前端IP地址和WiseLink DNS IP地址。	
动态前端IP地址	192.168.100.110～192.168.100.150		
管理网络网关地址	IP地址：10.110.100.1 掩码：255.255.255.0	用于OceanStor 9000节点与客户维护管理系统通信。	管理维护网络
OM管理IP	10.110.100.100	管理员可以使用该IP地址登录OceanStor 9000的管理系统。该IP地址为浮动IP地址。该地址可以浮动到集群内所有节点的前端网络接口。	

 Page 35 HUAWEI

- 当部署WiseLink特性时，需要规划动态前端IP地址和WiseLink DNS IP地址。其中动态前端IP地址需要在完成OceanStor 9000软件安装后，在OceanStor 9000管理控制台上配置。

IP地址规划样例（配置文件样例）

集群					
集群名称	N9000Cluster				
内部DNS	192.168.100.100	前端网络掩码	255.255.255.0	前端网络网关	192.168.100.1
后端网络掩码	255.255.255.0			数据冗余配比	+2
管理网络掩码	255.255.255.0	管理网络网关	10.110.100.1		
NTP Server地址	192.168.100.220	OM管理IP	192.168.100.200		
部署节点					
节点SN	210235G78PZ0D2000003	节点名称	N9000NODE1	IPMI IP	10.110.100.10
机架号	1	槽位号	1		
网口是否Bond	Yes				
Bond0的网口列表	B0-0, B0-1, B1-0, B1-1				
Bond1的网口列表	NIC0, NIC1				
Bond2的网口列表					
后端IP地址1	172.18.0.10	后端IP地址1的网口	bond0		
后端IP地址2		后端IP地址2的网口			
前端IP地址1	192.168.100.10	前端IP地址1的网口	bond1		
前端IP地址2		前端IP地址2的网口			
是否独立出管理网口	No	OM管理网口			

前端与后端接口配置为bond接口

默认情况下，不独立出管理网口，此时OM管理IP地址应在前端网络

HUAWEI

交换机配置说明

- 配置交换机管理IP
- 配置堆叠
- 配置SNMP
- 配置Telnet登录和Console登录

总结

- OceanStor 9000 10GE组网规划
- OceanStor 9000 IB组网规划
- OceanStor 9000 GE组网规划

思考题

- 请描述OceanStor 9000 10GE组网连接
- 请描述OceanStor 9000 IB组网连接
- 请描述OceanStor 9000 GE组网连接

习题

- 1、（单选题）一套OceanStor 9000系统中如果既有P节点又有C节点，那么总节点数至少为？

  A、3台；B、4台；C、5台；D、6台

- 2、OceanStor 9000支持以下哪几种组网？

  A、FE；B、GE；C、10GE；D、IB

- 1、D；2、BCD

Thank you

www.huawei.com

华为职业认证通过者权益

通过**任一项**华为职业认证，您即可在华为在线学习网站(*http://learning.huawei.com/cn*) 享有如下特权：

- **1、华为E-learning 课程学习**
  - **内容**：**所有华为职业认证E-Learning课程**，扩展您在其他技术领域的技术知识
  - **方式**：请提交您的 “华为账号” 和注册账号的 “email地址” 到 *Learning@huawei.com 申请权限。*
- **2、华为培训教材下载**
  - **内容：华为职业认证培训教材+华为产品技术培训教材**，覆盖企业网络、存储、安全等诸多领域
  - **方式**：登录*华为在线学习网站*，进入 “*华为培训->面授培训*” ，在具体课程页面即可下载教材。
- **3、华为在线公开课(LVC)优先参与**
  - **内容**：企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师授课，开班人数有限
  - **方式**：开班计划及参与方式请详见LVC排期：
    http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_16
- **4、学习工具 eNSP**
  - *eNSP (Enterprise Network Simulation Platform),* 是由华为提供的免费的、可扩展的、图形化网络仿真工具。主要对企业网路由器和交换机进行硬件模拟，完美呈现真实设备实景；同时也支持大型网络模拟，让大家在没有真实设备的情况下也能够进行实验测试。
- 另外，华为建立了知识分享平台 华为认证论坛。您可以在线与华为技术专家交流技术，与其他考生分享考试经验，一起学习华为产品技术。（http://support.huawei.com/ecommunity/bbs/list_2247.html ）